Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Speech Intelligibility Prediction for Hearing Aid Systems

Heidemann Andersen, Asger

DOI (link to publication from Publisher):
10.5278/vbn.phd.tech.00031

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Heidemann Andersen, A. (2017). *Speech Intelligibility Prediction for Hearing Aid Systems*. Aalborg Universitetsforlag. Ph.d.-serien for Det Tekniske Fakultet for IT og Design, Aalborg Universitet https://doi.org/10.5278/vbn.phd.tech.00031

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: May 05, 2023

SPEECH INTELLIGIBILITY PREDICTION FOR HEARING AID SYSTEMS

BY
ASGER HEIDEMANN ANDERSEN

DISSERTATION SUBMITTED 2017

AALBORG UNIVERSITY
DENMARK

Speech Intelligibility Prediction for Hearing Aid Systems

Ph.D. Dissertation

Asger Heidemann Andersen

Dissertation submitted August 31, 2017

Dissertation submitted: August 31, 2017

Industrial PhD Supervisors: Prof. Jesper Jensen
Oticon A/S and Aalborg University

Tekn. Dr. Jan Mark de Haan
Oticon A/S

University PhD Supervisor: Prof. Zheng-Hua Tan
Aalborg University

PhD committee: Associate Professor Flemming Christensen (Chairman)
Aalborg University

Dr. rer, nat.Thomas Brand
Oldenburg University

Reader Mike Brookes
Imperial College London

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Electronic Systems

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-050-0

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

© Copyright: Asger Heidemann Andersen, except where otherwise stated

Printed in Denmark by Rosendahls, 2017

About the Author

Asger Heidemann Andersen

Asger Heidemann Andersen received the B.Sc. degree in Electronics & IT and the M.Sc. degree in Wireless Communication (*cum laude*) from Aalborg University, Aalborg, Denmark, in 2012 and 2014, respectively. He is currently employed at Oticon A/S while pursuing the Ph.D. degree with the Signal and Information Processing section at Aalborg University under the supervision of Jesper Jensen, Jan Mark de Haan and Zheng-Hua Tan. His main research interests are prediction and measurement of speech intelligibility and speech enhancement with applications to hearing aids.

About the Author

Abstract

Hearing loss is a serious medical condition, which, primarily due to the aging of the population, affects increasingly many individuals. The typical approach to treatment involves the daily use of hearing aids, which amplify and enhance incoming sounds. Immense research efforts continuously strive to improve the performance of such hearing aids. While the performance of hearing aids can be measured in many ways, speech intelligibility is possibly the most important one. Designing hearing aids, which are good in this respect, is made difficult by the fact that speech intelligibility is cumbersome to measure. Typically, speech intelligibility has to be measured using time consuming and expensive listening experiments involving many listeners.

In this thesis we study an emerging alternative to listening experiments: Speech Intelligibility Prediction (SIP) algorithms. Such algorithms estimate the outcome of listening experiments, using audio recordings from the studied acoustical conditions. We propose and investigate several SIP algorithms extending the popular Short-Time Objective Intelligibility (STOI) measure. While these algorithms are applicable for a wide range of use-cases, we have investigated particular applications in the development of hearing aid systems.

Specifically, we propose 1) multiple *binaural* SIP algorithms, which can account for the advantage obtained in binaural listening environments, and 2) two *non-intrusive* SIP algorithms, which can predict intelligibility without requiring a clean speech reference signal. The proposed measures are shown to yield predictions which are accurate in comparison with competing algorithms, across a broad range of acoustical conditions including different configurations of noise types, non-linear processing, reverberation, and spatial source positions. In particular, using one of the proposed algorithms, it was possible to predict the result of a listening experiment, which had previously been carried out as part of product development within Oticon A/S. Thus, in principle, a subset of such experiments could be replaced by cheaper and less time consuming predictions. This underlines the potential of SIP as a highly valuable tool within the development of hearing aids.

Resumé

Høretab er en alvorlig lidelse, som – primært på grund af en aldrende befolkning – påvirker et sigende antal mennesker. Den typiske tilgang til behandling indebærer daglig anvendelse af høreapparater, som forstærker og forbedrer indkommende lyde. En enorm forskningsaktivitet stræber konstant efter at forbedre sådanne apparater. Ydelsen af et høreapparat kan måles på mange måder, hvoraf brugerens taleforståelse muligvis er den vigtigste. Det, at designe et høreapparat som forbedrer taleforståelsen markant, bliver besværliggjort af det faktum, at taleforståelse er svært at måle. Typisk måles taleforståelse igennem tidskrævende og dyre lytteforsøg, som involverer mange forsøgspersoner.

I denne afhandling studerer vi et nyere alternativ til lytteforsøg: Algoritmer til taleforståelighedsprædiktion (TFP). Sådanne algoritmer estimerer resultatet af lytteforsøg ved brug af lydoptagelser fra de undersøgte akustiske forhold. Vi foreslår og undersøger adskillige TFP-algoritmer, som alle er baseret på den populære Short-Time Objective Intelligibility (STOI) algoritme. Disse algoritmer kan anvendes til en lang række formål, men vi har specifikt undersøgt deres mulige anvendelse indenfor udvikling af høreapparater.

Helt konkret foreslår vi 1) flere *binaurale* TFP-algoritmer, som kan tage højde for den fordel der opnås i binaurale lytteforhold, og 2) to *ikke-invasive* TFP-algoritmer, som kan forudsige forståelighed uden at gøre anvendelse af et referencetalesignal. Det demonstreres, at de foreslåede algoritmer er mere nøjagtige end tidligere foreslåede algoritmer på tværs af en lang række akustiske miljøer bestående af forskellige kombinationer af støjtyper, ikke-lineære processeringstyper, rumklang og lydkildeplaceringer. Specifikt viser vi ved at anvende en af de foreslåede algoritmer, at det er muligt at forudsige resultatet af et tidligere lytteforsøg, udført som en del af produktudviklingen på Oticon A/S. Derved ses det, at det i princippet er muligt at udskifte en andel af de udførte lytteforsøg med billigere og mindre tidskrævende forudsigelser. Dette understreger potentialet for TFP-algoritmer som et værdifuldt værktøj indenfor udviklingen af høreapparater.

Contents

Abbreviations

ADFD	Akustiske Databaser For Dansk
AI	Articulation Index
ANSI	American National Standards Institute
ASR	Automatic Speech Recognition
beMBSTOI	better-ear MBSTOI
BFHN	Bottling Factory Hall Noise
BIF	Band Importance Function
BMLD	Binaural Masking Level Difference
BRIR	Binaural Room Impulse Response
B-sEPSM	Binaural sEPSM
BSIM	Binaural Speech Intelligibility Measure
BSTOI	Binaural STOI
CNN	Convolutional Neural Network
CSII	Coherence SII
DBSTOI	Deterministic BSTOI
DFT	Discrete Fourier Transformation
DRR	Direct to Reverberant Ratio
EC	Equalization Cancellation
ELA	Equal Local Azimuth
EPSM	Envelope Power Spectrum Model
ESII	Extended SII
ESTOI	Extended STOI
FF	Feed-Forward
FIR	Finite Impulse Response
fwSNRseg	frequency-weighted segmental SNR
HASPI	Hearing-Aid Speech Perception Index
HATS	Head And Torso Simulator
HINT	Hearing In Noise Test
HMM	Hidden Markov Model
HP	Highpass
HRTF	Head Related Transfer Function
IBM	Ideal Binary Mask
ILD	Interaural Level Difference

ISTS	International Speech Test Signal
ITD	Interaural Time Difference
ITFS	Ideal Time Frequency Segregation
LNOCV	Leave-N-Out Cross Validation
LP	Lowpass
MBSTOI	Modified BSTOI
MFCC	Mel Frequency Cepstral Coefficient
ModA	average modulation-spectrum area
mr-sEPSM	multi-resolution sEPSM
MVDR	Minimum Variance Distortionless Response
NI-STOI	Non-Intrusive STOI
RC	Relative Criterion
RIR	Room Impulse Response
RMSE	Root-Mean-Square Error
sEPSM	speech-based EPSM
SII	Speech Intelligibility Index
SIMI	Speech Intelligibility prediction based on Mutual Information
SIP	Speech Intelligibility Prediction
SNR	Signal to Noise Ratio
SRMR	Speech to Reverberation Modulation energy Ratio
SRT	Speech Reception Threshold
SSN	Speech Shaped Noise
stBSIM	short-time BSIM
STI	Speech Transmission Index
STMI	Spectro-Temporal Modulation Index
STOI	Short-Time Objective Intelligibility
TBM	Target Binary Mask
TF	Time-Frequency
VAD	Voice Activity Detector
WSTOI	Weighted STOI

List of Publications

The main body of this thesis consists of the following publications:

[A] A. H. Andersen, J. M. de Haan, Z.-H. Tan and J. Jensen, "A binaural short time objective intelligibility measure for noisy and enhanced speech," in *Proceedings of Interspeech*, pp. 2563–2567, 2015.

[B] A. H. Andersen, J. M. de Haan, Z.-H. Tan and J. Jensen, "A method for predicting the intelligibility of noisy and non-linearly enhanced binaural speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4995–4999, 2016.

[C] A. H. Andersen, E. Schoenmaker and S. van de Par, "Speech intelligibility prediction as a classification problem," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016.

[D] A. H. Andersen, J. M. de Haan, Z.-H. Tan and J. Jensen, "Predicting the intelligibility of noisy and nonlinearly processed binaural speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 24, No. 11, pp. 1908–1920, 2016.

[E] A. H. Andersen, J. M. de Haan, Z.-H. Tan and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5085–5089, 2017.

[F] A. H. Andersen, J. M. de Haan, Z.-H. Tan and J. Jensen, "On the use of band importance weighting in the short-time objective intelligibility measure," in *Proceedings of Interspeech*, pp. 2963–2967, 2017.

[G] A. H. Andersen, J. M. de Haan, Z.-H. Tan and J. Jensen, "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *submitted to Speech Communication*, 2017.

[H] A. H. Andersen, J. M. de Haan, Z.-H. Tan and J. Jensen, "Non-intrusive speech intelligibility prediction using convolutional neural networks," *submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

Furthermore, the following academic presentations have been given during the project:

[I] A. H. Andersen, J. M. de Haan, Z.-H. Tan and J. Jensen, "Predicting the intelligibility of noisy and enhanced binaural speech," poster presentation at the *Speech in Noise Workshop*, Groningen, Netherlands, 2016.

[J] A. H. Andersen, J. M. de Haan, Z.-H. Tan and J. Jensen, "Performance evaluation of the short-time objective intelligibility measure with different band importance functions," poster presentation at the *Speech in Noise Workshop*, Oldenburg, Germany, 2017.

[K] A. H. Andersen, "Binaural Intelligibility Prediction for Noisy and Nonlinearly Processed Speech," invited talk at the *Speech in Noise Workshop*, Oldenburg, Germany, 2017.

The following patent applications have been filed in relation to the project:

[L] J. Jensen, A. H. Andersen and J. M. de Haan, "A hearing system comprising a binaural speech intelligibility predictor," European patent application, EP16152967.2, 2015.

[M] J. Jensen, A. H. Andersen and J. M. de Haan, "A monaural speech intelligibility predictor unit, a hearing aid and a binaural hearing system," European patent application, EP17153174.2, 2016.

[N] J. Jensen, J. M. de Haan and A. H. Andersen, "A monaural intrusive speech intelligibility predictor unit, a hearing aid and a binaural hearing aid system," European patent application, EP16157993.3, 2016.

[O] A. H. Andersen, J. M. de Haan, Z.-H. Tan, J. Jensen and M. S. Pedersen, "A method for predicting the intelligibility of noisy and/or enhanced speech and a binaural hearing system," European patent application, EP16160309.7, 2016.

Preface

This thesis documents the scientific work carried out as part of the industrial Ph.D. project "*Speech Intelligibility Prediction for Hearing Aid Systems*". The project was carried out as a joint venture between Oticon A/S (Smørum, Denmark) and Aalborg University (Aalborg, Denmark), supported by the Danish Innovation Foundation[1] and the Oticon Foundation[2]. The work was carried out between September 2014 and August 2017, and mainly took place within the Audiological Design and Signal Processing group at Oticon A/S, as well as in the Signal and Information Processing section at Aalborg University. Furthermore, parts of the work were carried out in the Acoustics Group at the University of Oldenburg, during a four-month secondment.

The thesis is structured in two parts: a general introduction and a collection of scientific papers. The introduction contains a broad introduction to the field of speech intelligibility prediction, and furthermore summarizes the contributions of the Ph.D. project. The introduction is followed by a collection of eight papers, published through different outlets.

I am thankful to many friends and colleagues who have supported me in carrying out this project. First and foremost, I very grateful to my three highly skilled and motivated supervisors, Jesper Jensen, Jan Mark de Haan, and Zheng-Hua Tan, who have generously spent countless hours to guide me. In particular, I am deeply indebted to Jesper Jensen who has always been available for lengthy and inspiring discussions, and who has tirelessly read, commented, re-read, and re-commented endless drafts of questionable literary quality. I also wish to thank Steven van de Par, Esther Schoenmaker, and the rest of the Acoustics Group at the University of Oldenburg, for allowing me to have an inspiring stay. Similarly, I am thankful to colleagues in and around the Audiological Design and Signal Processing (ADSP) group at Oticon, and in the Signal and Information Processing section at Aalborg University. In particular, I wish to thank Adam Kuklasiński and Mojtaba Farmani who have supported me immensely through their friendship, and through the sharing of countless tips and tricks from their recent experiences

[1] `https://innovationsfonden.dk/en/application/erhvervsphd`.

[2] `http://www.oticonfoundation.com/`.

of thesis-writing. Lastly, I wish to thank my family for supporting me and standing by my side through the entire process of this work.

Asger Heidemann Andersen
Copenhagen, August 31, 2017

Part I

Introduction

Introduction

1 The Challenges of Speech Communication

The use of sound to communicate information is common throughout the animal kingdom. In humans, this ability has evolved to an outstanding level of sophistication. Our ability to produce and understand speech allows us to communicate highly complex and varied information such as knowledge, intents, opinions and emotions.

The question of how humans have come to evolve speech and language remains fiercely debated and largely unanswered [51, 52]. Nevertheless, one cannot dispute the importance of verbal communication in the current state of human living.

Because of the importance of speech, the technical sciences have always striven to overcome the natural boundaries of vocal communication: telephones allow us to have conversations across long distances, microphones, loudspeakers and electronic data storage allow us to store and replay spoken words, and hearing aids allow us to understand speech in spite of dysfunctional hearing. Nowadays, devices for speech processing and transmission are equipped with powerful noise reduction and speech enhancement algorithms, to make them function in ever noisier environments. This is made possible by advances in available computational power, battery technology, and advanced signal processing.

The increasing number of possibilities within speech technology make it continuously more important to understand as much as possible about speech communication. By now, the physiology and basic operation of the speech production facility – the vocal tract – and the hearing facility – the ear – are rather well understood. However, much less is known about the underlying operation of the human brain [73, 115].

1.1 A Model of Speech Communication

For the purposes of this thesis, we consider the simple model of speech communication shown in Fig. 1. This consists of 1) a talker, using her vocal system

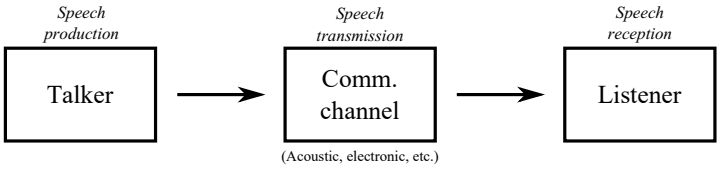


Fig. 1: A general diagram showing how information is transmitted from a talker (producing speech), via a communication channel, to a listener (receiving and interpreting speech). In the simple case of a face-to-face conversation, the communication channel is purely acoustical, and adds noise and reverberation to the transmitted speech signal. In more complex scenarios, the channel could include microphones, loudspeakers, distortion, processing, storage, radio transmission, etc.

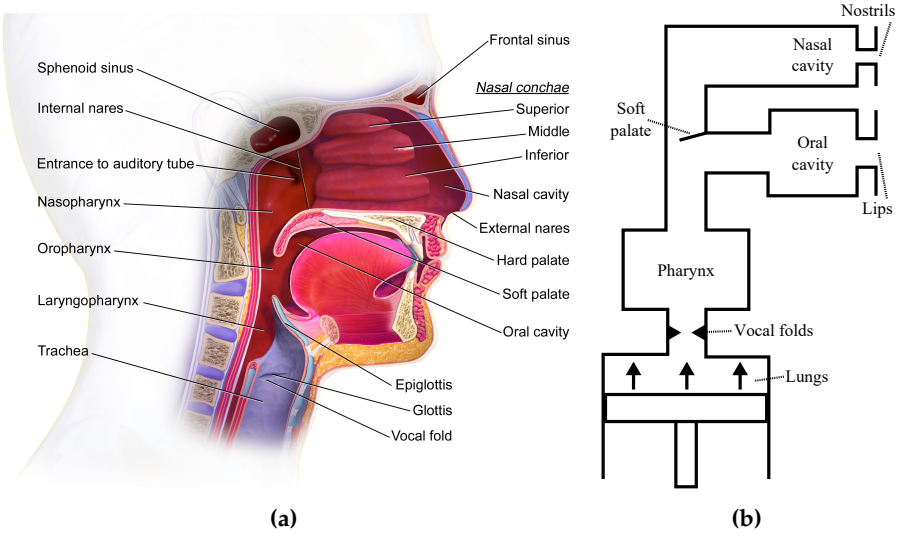


Fig. 2: (a) An anatomical drawing of the vocal tract. Adapted from [9]. **(b)** A functional diagram of the operation of the vocal tract. Inspired by a similar figure in [23].

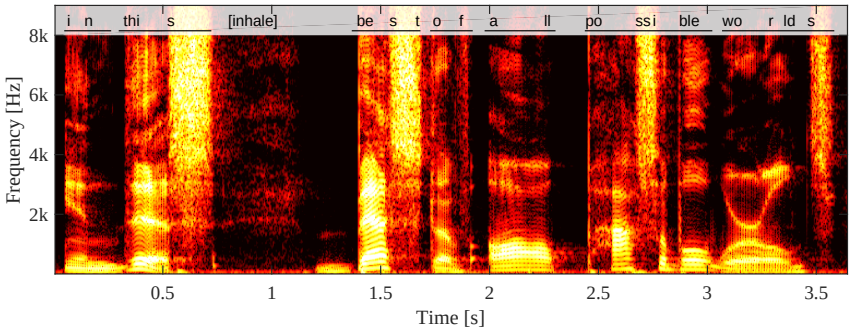


Fig. 3: A spectrogram of a male saying "*in this, best of all possible worlds*".

to generate speech, 2) a communication channel which carries the speech signal to the listener, and 3) a listener, who uses his ear to hear the output of the channel, and attempts to decode the speech. This simple model is useful, as it captures the essentials of the speech communication process. It should be noted, however, that the model has limitations. For example, it assumes that information is transmitted only by speech (and not e.g. by body language). It also assumes communication to be entirely a one-way process. However, it fits the scope of this thesis very well, as we study intelligibility mainly in scenarios which comply with these limitations. The communication channel may be either purely acoustical or contain elements of electronic processing (e.g. a telephone or a hearing aid system). We treat the three elements of the model separately in the three following sections.

Speech Production

The first block in Fig. 1 represents the formulation and utterance of a sentence by the talker. The formulation of sentences, and the more general representation of speech and language in the brain, are studied in such fields as linguistics [121] and neurolinguistics [73], but remain beyond the scope of this thesis. Instead, we focus on the mechanical production of speech sounds by the speech production facility, i.e. the vocal tract and lungs. We do so, mainly to obtain an intuition about the origin and the properties of the speech signal.

Fig. 2a shows an anatomical drawing of the vocal tract, and Fig. 2b shows a corresponding functional diagram. The bottom of Fig. 2a depicts the larynx (in blue), containing the vocal folds and the glottis. To produce voiced sounds, air is forced out of the lungs, while the vocal folds close off to restrict the flow of air. This causes a build-up of air pressure, which periodically forces the vocal folds apart to release impulsive bursts of air at a rate of 60–300 Hz [23, 71]. These bursts proceed through the pharynx, the cavity located behind the oral and nasal cavities, and finally escape through the mouth and nose. Throughout this process, muscular activity in the larynx, pharynx, and oral and nasal cavities results in a time-varying filtering of the excitation signal. This gives rise to time-varying spectral peaks and troughs in the resulting signal, which are used to encode information [23, 139].

The above described process is responsible for the generation of vowel sounds. Speech also includes a range of other sounds, which may be generated with or without the active use of the vocal cords [23]. These include consonant sounds such as fricatives (e.g. [s] and [f]), which are produced by creating a narrow constriction at some point along the vocal tract, resulting in a turbulent airflow, emitting sound with noise-like characteristics [139]. Another characteristic speech sound is the stop (e.g. [p], [t] and [k]), where the flow of air is either suddenly blocked by a constriction, or released by the sudden removal of a constriction [139].

For a more detailed treatment of sounds involved in spoken language, we refer to [91]. For a more detailed review of both the anatomy and the acoustical properties of the vocal tract, we refer to [139].

Features of the resulting speech signal can be visualized effectively by use of a spectrogram, as done in Fig. 3. Harmonics, resulting from the opening and closing of the vocal folds, are seen as horizontal stripes in voiced periods; especially clearly at around 2.0–2.3s into the signal. Time-varying resonances in the vocal tract create changing spectral peaks, called formants [116]. Three formants are clearly visible 0.4s into the signal on Fig. 3. Another vivid feature seen in Fig. 3 is the contrast between the voiced segments (e.g. 1.5s into the signal), with much low frequency energy, and fricatives (*s* and *f* sounds, e.g. 1.6s into the signal) which typically have less powerful, noise-like, characteristics and more high-frequency energy.

Speech Transmission

Once the sound signal has exited the mouth and nose of the talker, it travels to the ear of the listener through an acoustic or partially acoustic communication channel as indicated in Fig. 1. During the transmission process, the signal may be degraded by a large number of factors. In the simplest case, the talker and the listener are located close to one-another, and the sound can travel along a purely acoustical path. In this case, the speech signal is mainly impacted by environmental noise and, perhaps, reverberation.

In a slightly more complex scenario, the talker and listener may communicate through a pair of mobile phones. In this situation, the speech of the talker is converted into one or more electrical signals by the microphone(s) located close to the mouth of the talker. This signal is passed through a complex signal processing chain involving analog-to-digital conversion, speech enhancement and coding, wireless transmission across multiple wireless channels, and digital-to-analog conversion. The resulting electronic signal is turned back into an acoustic signal near one ear of the listener. Environmental noise and reverberation may obviously still play a role in this situation. All parts of this process should be considered part of the communication channel in Fig. 1.

A third example, of particular relevance to this thesis, is that in which the talker and the listener are near to one-another and can thus communicate directly by acoustic speech, but in which the listener is wearing hearing aids to compensate for a hearing loss. In this case, the communication channel includes a direct acoustic path, which circumvents the hearing aid, e.g. sound passing through a ventilation channel in the hearing aid [32], and another path which includes the signal processing of the hearing aid and its microphone/loudspeaker characteristics. We remark that the hearing aid is considered part of the communication channel, while the hearing loss is con-

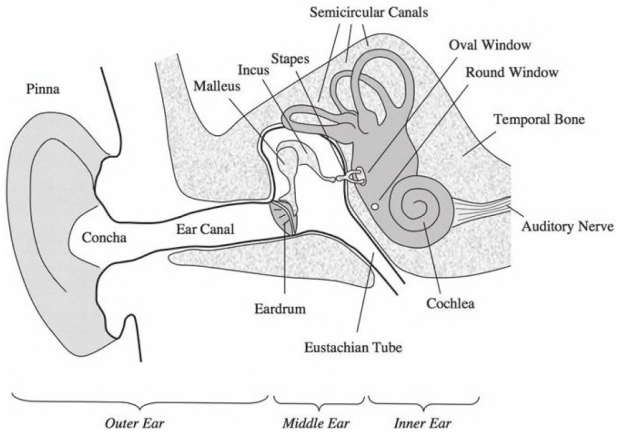


Fig. 4: A cross section of the human ear. Figure reproduced from [116] with permission.

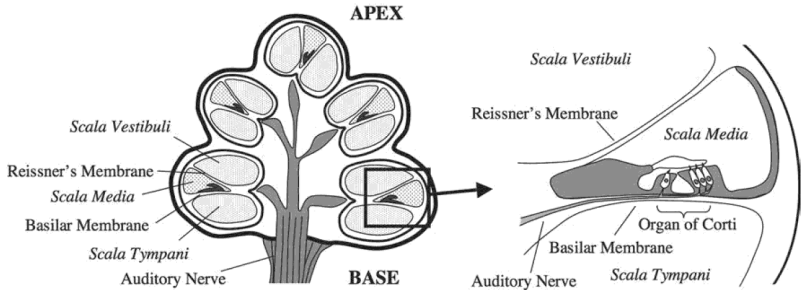


Fig. 5: A cross section of the cochlea. Figure reproduced from [116] with permission.

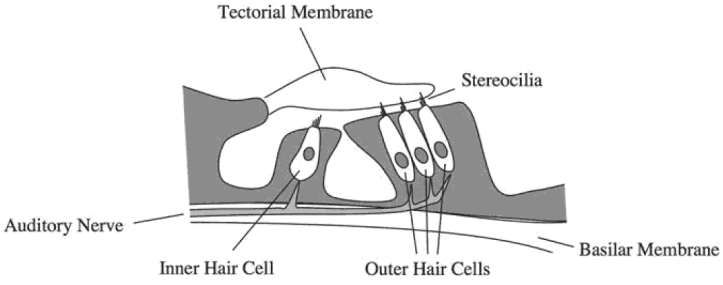


Fig. 6: A cross section of the basilar membrane, showing the organ of Corti. Figure reproduced from [116] with permission.

sidered part of the listener. This is an example in which the effect of the channel is partially to enhance the speech signal.

Speech Reception

We now continue to the third block in Fig. 1 in which the acoustic speech signal enters the ear of the listener. Fig. 4 shows an overview of the human ear. As a sound wave impinges on the ear, it encounters the pinna (the external, visible, part of the ear), which causes multiple reflections to enter the ear canal. The reflections from the pinna, which are dependent on the direction of arrival, may aid the listener in determining the position of the talker [8]. The ear canal is terminated by the eardrum, which separates the outer ear and the middle ear. The eardrum converts acoustic waves into mechanical waves, which are conducted to the inner ear by the ossicles, the three bones located in the middle ear: the malleus, the incus, and the stapes. The inner ear consists of the spiral-shaped cochlea, and a number of semicircular canals, which are involved in our sense of balance [137]. The cochlea converts mechanical vibrations, entering from the stapes, into neural activity, exiting through the auditory nerve [116]. Fig. 5 shows a cross section of the cochlea. This reveals that the cochlea takes the form of a coiled-up tube, separated into three fluid-filled chambers: scala vestibuli, which terminates in the oval window connected to the stapes, scala tympani which terminates in the round window, and scala media. The membrane between scala tympani and scala media is known as the basilar membrane. When mechanical oscillations enter the cochlea via the stapes, the oscillations propagate along the basilar membrane from base to apex. The basilar membrane is mechanically tuned to resonate at higher frequencies towards the base and at lower frequencies towards the apex [115]. Thus, the basilar membrane effectively performs a mechanical frequency analysis of the sounds entering the ear. For this reason, the cochlea is often modelled as a bank of bandpass filters [53, 107]. On the basilar membrane lies the organ of Corti, as shown in Fig. 5 and Fig. 6. The organ of Corti consists of a row of inner hair cells and multiple rows of outer hair cells [116]. Each of the hair cells are covered by a number of stereocilla (i.e. hairs). The hair cells are covered by the tectorial membrane. When the basilar membrane vibrates, the tectorial membrane causes the stereocilla to displace. In the inner hair cells, this triggers a chemical reaction which results in a signal being sent along the auditory nerve. The outer hair cells are believed not to contribute directly to the signals sent along the auditory nerve, but rather to actively amplify weak vibrations on the basilar membrane, making them more detectable by the inner hair cells [107]. As a consequence, the outer hair cells appear to increase the sensitivity and frequency resolution of the ear [107, 116]. The signals transmitted across the auditory nerve are processed in a complex network of neuronal structures, known as the audi-

tory pathway, concluding in the primary auditory cortex [116]. For a more detailed overview of the physiology of the ear and the related parts of the brain, we refer to [115].

The healthy auditory system is remarkable in its ability to sense acoustic stimuli across a wide range of frequencies (around 20 Hz to 20 kHz) and across an extremely large dynamic range (around 120 dB). However, a number of rather common medical conditions can diminish the performance of the ear. Hearing losses are typically classified as either conductive or sensorineural [116]. A conductive hearing loss is a condition in which sound is somehow attenuated before reaching the cochlea. This could be due to e.g. earwax, perforation of the eardrum, or a buildup of fluid in the middle ear [116]. A sensorineural hearing loss is a condition in which the functioning of the cochlea or the auditory nerve has been damaged. Commonly, the outer hair cells are damaged either by exposure to loud sound, old age, or genetic factors. However, recent research has revealed the existence of types of sensorineural hearing loss which cannot be explained as damage to the cochlea, and are therefore likely to be caused by damage further along the auditory nerve [56, 64]. Sensorineural hearing losses are the most common type of hearing loss, and occur especially frequently among the elderly [116]. According to estimates by the World Health Organization (WHO), 360 million people suffer from disabling hearing loss globally[3].

1.2 Speech Communication in Noisy Environments

Humans can use speech to communicate effectively even in conditions with strong adverse factors such as noise, interfering speakers, reverberation, distortion, or hearing impairment. This ability was most famously summarized in 1953 by E. Colin Cherry's definition of the "*cocktail party problem*" in which a listener must "*recognize what one person is saying when others are speaking at the same time*" [22]. This ability of humans, to selectively attend to one talker, while somehow filtering away meaningful speech from dozens of other nearby talkers, has baffled researchers ever since [14]. Many researchers would likely argue that the arcane properties of human speech communication alone justify its scientific study. However, another important justification comes from the fact that speech communication often becomes uncomfortable or entirely impossible, when conditions are simply too adverse. Today, an abundance of technological tools exist, which could potentially ease speech communication in such situations. The successful application of these tools is, however, often dependent on an understanding of what factors make speech communication more or less difficult.

The study of speech communication in difficult environments, can be approached with several foci, such as listening effort [128], speech qual-

[3] `http://www.who.int/pbd/deafness/estimates/en/`. Accessed on 28/08-2017.

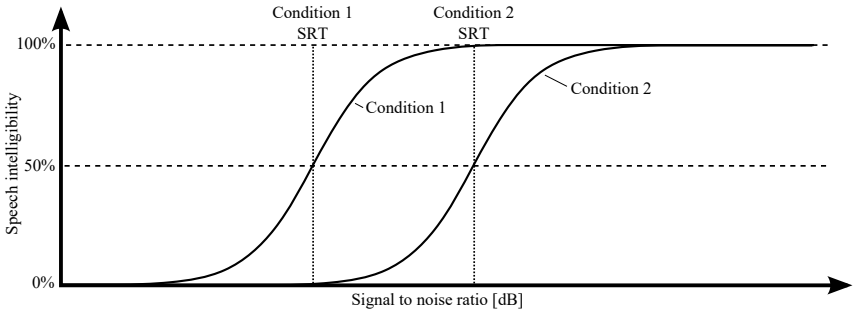


Fig. 7: An illustration of how intelligibility could depend on SNR in two different listening conditions. The SRTs of the two conditions are marked by vertical dotted lines.

ity [100, 153], or speech intelligibility [104]. While they are all relevant, depending on the context, we have focused entirely on speech intelligibility in the work underlying this thesis, believing this to be the most important measure of successful speech communication.

In this section, we summarize how speech intelligibility is scientifically studied, and how different factors have been found to influence speech intelligibility.

Measuring Speech Intelligibility

In order to investigate the functioning of human speech communication, researchers typically make use of experiments where intelligibility is measured in difficult, but highly controlled, conditions. In such an experiment, prerecorded sentences (or other tokens of speech, such as words, syllables, or phonemes) are played back to a test subject in a sound studio or a sound isolated booth. The sound may be played back either via a loudspeaker system or via headphones. The subject attempts to understand and repeat the presented sentences. Meanwhile, the experimenter keeps track of the number of syllables, words, or sentences correctly repeated by the subject. The task is made difficult by adding noise or otherwise degrading the speech signal heard by the listener. The addition of noise has the advantage that it allows for adjusting the intelligibility by changing the Signal to Noise Ratio (SNR), i.e. the ratio of speech power to noise power. Plotting speech intelligibility as a function of SNR yields a characteristic S-shaped curve which approaches 0% for low SNRs and approaches 100% for high SNRs (see Fig. 7). The intelligibility at a particular SNR is dependent on the noise type, and a range of other characteristics of the presented acoustical condition. Speech intelligibility is typically measured in one of two ways: 1) intelligibility is measured at a particular SNR and reported as a fraction of correctly repeated syllables, words, or sentences, or 2) the SNR is adaptively changed from sentence

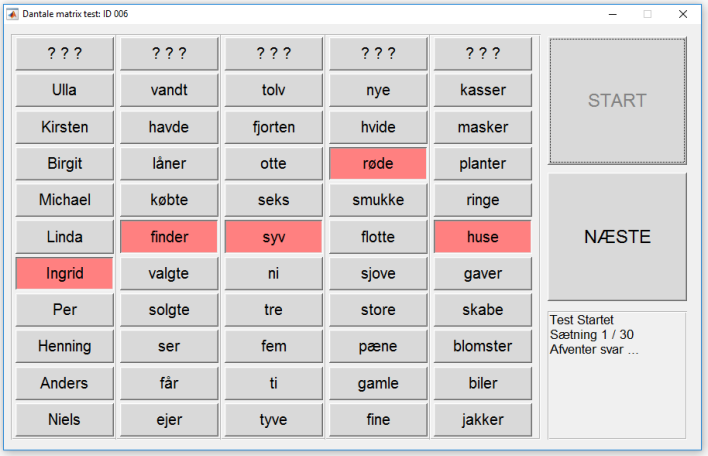


Fig. 8: A software tool for collecting listening test results with the Danish Dantale II speech material. The test subject is presented with a sentence, and hereafter attempts to build the sentence by selecting one word from each column of buttons.

to sentence, such as to estimate the SNR at which the subject can repeat a fixed fraction of syllables, words or sentences [96, 97, 117]. In the latter case, the SNR is most often adapted towards an intelligibility of 50%, as this is the point which can be estimated most accurately [10, 142]. We refer to the SNR at which intelligibility is 50% as the Speech Reception Threshold (SRT).

Several types of speech corpora, consisting of sentences, words [40], phonemes [103], or other speech tokens, have been devised for use in controlled listening experiments. In this work, we have mainly considered sentence corpora. One typically distinguishes between *open set* and *closed set* sentence corpora. Open set corpora contain recordings of separate, possibly semantically meaningful, sentences with no direct limitation in vocabulary, e.g. common everyday sentences. Examples of such corpora are the Harvard sentences [72] and the Hearing In Noise Test (HINT) sentences [110]. Closed set corpora contain sentences which are generated from a small collection of words. The most common type of closed set corpora are the matrix sentence tests, which have been developed for a considerable number of languages [60, 62, 66, 86, 111, 119, 146–149, 152, 154]. These consist of five-word sentences generated by randomly sampling words from five lists of ten words [60]. An advantage of matrix sentence tests is that subjects can enter perceived sentences in a graphical user interface, as illustrated in Fig. 8, avoiding the need for the presence of an experimenter [112]. The Danish version of the matrix sentences, the Dantale II corpus [149], has been used for the majority of the listening experiments carried out as part of the work underlying this thesis.

Speech intelligibility can be adversely affected by factors from all three blocks in Fig. 1. A considerable amount of research has been directed towards establishing the impact, on intelligibility, of many such factors. The remainder of this section is used to review a selection of these.

The Impact of the Talker on Speech Intelligibility

The talker should be able to produce an understandable sentence, in a language which is familiar to the listener, in order for speech communication to be possible. The ease of communication can vary greatly, depending on the language proficiency of the listener, and the accent of the talker. However, one could also speculate that some languages are intrinsically more intelligible than others. While this is extremely difficult to investigate in controlled conditions, studies have shown that nearly identical listening experiments with native listeners yield different results when carried out in different languages [63]. Another notable effect, is that humans tend to adapt their speaking style to the environment by, for example, speaking more loudly in the presence of noise [92].

The Impact of the Communication Channel on Speech Intelligibility

The most variable and commonly experienced obstacles to speech communication are located in the communication channel between the talker and the listener. The most common of these is probably environmental noise, but several other properties of the channel impact speech intelligibility. We shortly summarize the most important ones:

- **Noise:** Noise can have many distinctive sources such as machinery, vehicles, ventilation systems, weather, or other human talkers. Noise types are often characterized according to their spectral properties (and especially their relative spectral similarity or dissimilarity to speech), as well as their temporal properties (stationary or fluctuating). These properties have been found to have important influences on speech intelligibility:
  - **Spectral attributes:** Speech is masked most effectively by noise types which contain energy at the same frequencies as speech [55, 104]. This can be attributed partly to the mechanical frequency analysis carried out by the cochlea, as this allows the auditory system to, at least partly, ignore noise components if they do not spectrally overlap with the signal of interest. Listening experiments often make use of Speech Shaped Noise (SSN), i.e. stationary Gaussian noise which has been filtered to have the same long-term spectrum as speech.

  - **Temporal attributes:** Speech intelligibility is generally high, for normal-hearing listeners, in interrupted or strongly fluctuating noise types, as compared to stationary noise [5, 30, 35, 36, 48–50, 59, 104, 105, 109]. This suggests that humans are capable of "listening in the gaps", i.e. to collect speech information quickly in short instants with high SNR [25, 49, 69, 70, 114].
  - **Speech-like interferers:** In the presence of one or more interfering talkers, the listener may be confused by the similarity of target speech and interfering speech. This, in turn, can lead to lower intelligibility [14, 107]. This is referred to as informational masking [107]. Speech intelligibility is particularly low when masked by "babble noise" consisting of multiple interfering speakers [133]. This is likely because babble noise is both spectrally similar to speech and contains modulations similar to speech, while also being sufficiently stationary to not allow for gap listening [133].

- **Reverberation:** This occurs because sound is reflected by walls and other surfaces, leading to temporal smearing of both speech and noise signals. Under some circumstances, reverberation can significantly degrade speech intelligibility [16, 18, 37, 106, 108].

- **Distortion:** In conditions where the communication channel includes a hearing aid, a telecommunications device, or another electronic device, electronically introduced distortion (e.g. harmonic distortion caused by non-linear device characteristics) may have a negative effect on speech communication [77, 83].

- **Spatial aspects:** The detrimental influence of noise and reverberation is to a great extent determined by the relative positions of the talker, the listener, and the noise sources. Firstly, this is obviously due to the fact that sound is attenuated with distance. If the talker and the lister are very close, while any noise sources are far away from both of them, the influence of noise is typically little. Secondly, humans have, not one, but two ears, located on opposite sides of the head. Due to the acoustical influence of the head, considerably different sound signals may enter the two ears. Many studies have shown that the human auditory system is very good at combining information from the two ears such as to improve intelligibility [15, 17, 18, 33, 36, 113, 118]. The advantage gained by having two ears is termed the binaural advantage. Such an advantage occurs in conditions where the target talker and the interfering noise source(s) are spatially separated. The binaural advantage can be decomposed into two separate components [18, 155]:

  1. Because of the acoustics of the human head, the SNR may be better

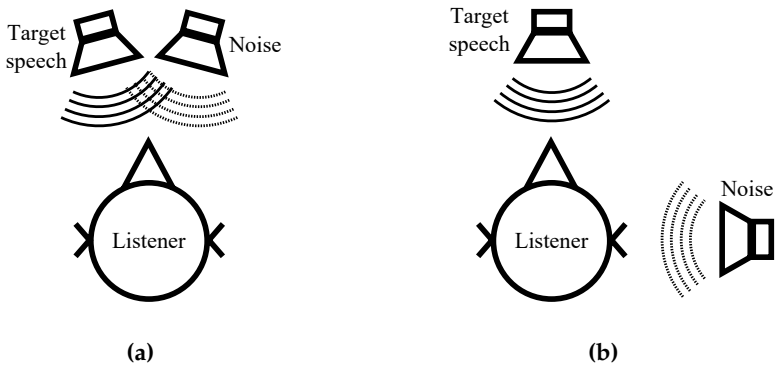


Fig. 9: Two acoustical environments with different amounts of binaural advantage. **(a)** Little binaural advantage can be obtained, as speech and noise comes from almost the same direction. **(b)** A considerable degree of binaural advantage can be obtained, due to both better-ear listening (the left ear is presented with a high SNR, as the head attenuates the noise) and binaural unmasking (the noise arrives at the left ear later than at the right ear).

at one ear than at the other. The auditory system can listen selectively to the ear with the higher SNR [15, 18], and the resulting improvement in intelligibility is referred to as a *better-ear* advantage.

2. Because of the distance between the two ears, acoustic waves arrive at the two ears with an interaural time delay, which depends on the direction of arrival [107]. This allows the auditory system to distinguish spatially separated sources, and leads to improved intelligibility [15, 18]. This is referred to as an advantage due to *binaural unmasking*.

Fig. 9 illustrates two different listening conditions: one with little or no binaural advantage (Fig. 9a), and one with a considerable degree of binaural advantage due to both better-ear advantage and binaural unmasking (Fig. 9b). Binaural advantage can lead to improvements in SRTs of as much as 12-14 dB [15, 18, 113, 118]. An extensive survey of binaural advantage is provided in [18].

The Impact of the Listener on Speech Intelligibility

Lastly, we shall mention the effect of hearing loss. Assuming the talker and the listener to be natives in the same language, this is perhaps the main influence of the listener, on speech intelligibility. The most common type of hearing loss, the sensorineural one, typically lowers both the sensitivity and the frequency resolution of the ear [107]. This, in turn, lowers speech

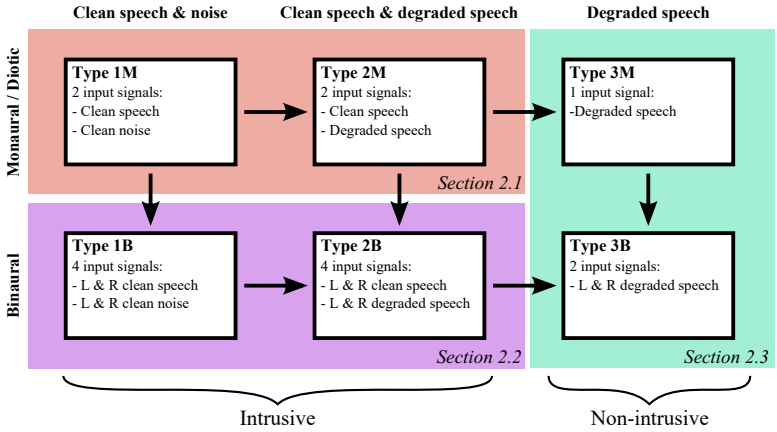


Fig. 10: A "taxonomy" of SIP algorithms, according to their respective input signal domains. The upper row contains algorithms which predict monaural or diotic intelligibility, while the bottom row contains methods which predict binaural intelligibility. The left column contains predictors which assume knowledge of speech and noise in separation. The middle column contains methods which assume knowledge of clean speech and degraded speech. The right column contains predictors which assume only access to degraded speech. An arrow from one type to another indicates that the input signal domain of the latter is a superset of the input signal domain of the former (e.g. a type 2M algorithm can make predictions for any input signals that a type 1M predictor can make predictions for, and more).

intelligibility in all situations. Also, hearing impairment degrades the ability to listen in the gaps, and thus lowers intelligibility severely in conditions with fluctuating noise types, including interfering talkers [5, 18, 30, 36, 50, 59]. The head shadow, which is an important factor in binaural advantage, occurs mainly at higher frequencies. Because hearing impairment is typically most severe at high frequencies, hearing impaired listeners suffer from reduced better-ear advantage [18, 36]. However, the ability to benefit from binaural unmasking often remains intact [17, 18].

2 Speech Intelligibility Prediction

Because speech processing, in general, and enhancement, in particular, are very important applications of signal processing [100], researchers have long striven to gain an understanding of what factors make speech intelligible, and what factors prevent it from being so. This has led to a broad range of empirical results, such as those discussed in Section 1.2. However, much effort has also been directed towards the development of predictive models, which can account for such empirical results.

In this thesis we study algorithms and models which predict intelligibility, based on recordings of noisy or otherwise degraded speech. The algorithms

typically output a single number, representing a "scoring" of intelligibility on some arbitrary scale. Through a calibration process, such a number can be converted into either a prediction of the fraction of correctly understood words, or a prediction of the SRT.

Early work on modeling speech intelligibility was carried out at Bell Laboratories in the 1920s and onwards [54, 55]. Since then, a large number of SIP algorithms have been proposed, for predicting how intelligibility is impacted by a range of factors such as noise level, noise type, processing, distortion, reverberation, noise source position, etc.

An important primary characteristic of SIP algorithms is the required type of input signals. Fig. 10 provides a "taxonomy" of methods, according to this characteristic. The figure distinguishes SIP algorithms according to whether they predict monaural/diotic[4] intelligibility (and thus require only input signals for one ear), or binaural intelligibility (and thus require signals from both ears of the listener). We refer to these as type M and type B algorithms, respectively. Furthermore, Fig. 10 distinguishes algorithms according to the signals required for making predictions. Type 1 algorithms predict the intelligibility of speech contaminated by additive noise, through knowledge of the clean speech and noise in separation. Importantly, such algorithms are fairly limited, in that they do not predict intelligibility for non-linearly processed noisy speech. This is because speech and noise are inseparable after the combination of the two has been non-linearly processed. Type 2 algorithms overcome this problem by requiring knowledge of clean speech (before potential non-linear processing), and degraded speech, which may be both noisy and non-linearly processed. The speech intelligibility of the degraded signal is, thus, predicted by comparing it to the clean signal. Type 1 and 2 algorithms are collectively referred to as *intrusive* SIP algorithms. Type 3 algorithms make predictions based on only the degraded speech signal itself. These are referred to as *non-intrusive* SIP algorithms. As shown with arrows on Fig. 10, there is a clear hierarchy among the six types of algorithms. A type 3 algorithm can be used as a type 2 predictor; simply by disregarding the clean signal. A type 2 algorithm can be used as a type 1 algorithm; because the degraded signal, in this case, is given by the sum of speech and noise. At the same time, a binaural algorithm can be used to make monaural or diotic predictions, either by letting the input signal to one ear equal zero, or by using the same input signals for both ears. In this sense, the type 3B algorithm is the most general one, as it can be used to make predictions for any type of input signals covered by Fig. 10. Conversely, the type 3 algorithms are given much less information, and it is therefore reasonable to expect them to make less accurate predictions than the type 1 and 2 algorithms. Similarly, type 2

[4] "Monaural" refers to the presentation of sound to one ear only, while "diotic" refers to the presentation of identical sound to both ears. Oppositely, "binaural" or "dichotic" refers the presentation of different sound to each ear.

algorithms should be expected to give predictions which are less accurate than those of type 1 algorithms.

The remainder of this section reviews the literature of SIP algorithms. We do not by any means attempt to exhaustively cover all existing methods, but rather focus on methods which are relevant to the present work. As indicated by Fig. 10, Section 2.1 covers monaural/diotic intrusive algorithms, Section 2.2 covers binaural intrusive algorithms, while Section 2.3 covers non-intrusive algorithms.

2.1 Intrusive Prediction of Monaural or Diotic Intelligibility

Monaural intrusive SIP has been the starting point for SIP as it is arguably the simplest case. Research on intrusive monaural SIP can be traced back to Bell Laboratories in the 1920s [54]. Two decades later, this work culminated in the proposal of the Articulation Index (AI) [55], which was later standardized [1]. Initially, a graphical procedure for pen-and-paper computation of the AI was used [88]. The advent of modern computing, however, made such approaches obsolete. This resulted in the Speech Intelligibility Index (SII) [2], an updated version of the AI, which is better suited for evaluation on a computer. The core assumption behind both the AI and SII, is that the speech spectrum can be separated into narrow frequency bands, and that each of these bands contribute independently to intelligibility [55]. Thereby, the intelligibility of speech can be summarized by an index, computed as follows [2]:

$$\mathrm{SII} = \sum_{i=1}^{n} I_i A_i, \tag{1}$$

where n is the number of frequency bands, I_i is the relative importance of the i'th band, and A_i is a measure of the speech fidelity in the i'th band. The values of I_i, collectively known as the Band Importance Function (BIF), are positive and sum to one. The BIF is determined from the results of a listening experiment, using a graphical procedure described in [55]. A more recent method for determining BIFs is based on minimizing prediction errors for a dataset of measured intelligibility [82]. The values of $A_i \in [0, 1]$ are determined from the spectra of speech and noise in the particular condition in question. The procedure for computing A_i takes a number of masking effects into account, but is mainly dependent on the band-wise SNR (see [2] for details on this). The resulting index is designed to have a monotonic relationship with measured intelligibility, and can be translated into predictions of measured intelligibility, given additional information about the used speech material [2]. The SII is able to account for individual hearing loss[5] by includ-

[5]By "modeling individual hearing loss" we mean that an algorithm is able to adapt its predictions to subject-specific performance, based on individual psychophysical measurements such as the audiogram.

ing a listener-dependent hearing threshold [2]. With reference to Fig. 10, both the AI and SII are SIP algorithms of type 1M, as they require knowledge of speech and noise spectra in separation.

Type 1M SIP Algorithms

The AI and SII constitute an important starting point in the field of SIP, and many further developments in the field can be seen as attempts to extend the domain in which they can predict intelligibility accurately. Specifically, we mention the following areas in which the AI and SII are known to fall short:

- The SII and AI rely on long-term frequency spectra of speech and noise (i.e. estimated for a period of several seconds). This prevents them from adequately accounting for short-term fluctuations in the noise. In practice, it is known that normal-hearing humans are very good at exploiting short pauses in the interferer signal to collect information from the speech signal (as discussed in Section 1.2). Thus, the AI and SII do not accurately reflect intelligibility in conditions with non-stationary interferers [124, 125].

- The SII and AI rely on knowledge of the speech and interferer spectra in separation. If the noisy speech has been non-linearly processed, these spectra are not well-defined. Thus, in such conditions, the AI and SII are unable to generate predictions at all.

- The AI and SII rely on the speech and interferer spectra presented to either one or both ears of the listener, but do not account for the possibility that different signals are presented to the left and right ears of the listener (as is the case in most real-world environment). Hence, the AI and SII are unable to account for binaural advantage.

The above shortcomings are highly relevant for practical applications of SIP, as real-world noise is often not stationary (e.g. interfering speech), nor monaural or diotic. Furthermore, an increasingly popular use-case for SIP algorithms is the evaluation of speech enhancement algorithms, which are often non-linear [100] and may also take binaural aspects into consideration [26, 34, 57, 61, 89, 90, 102].

A rather simple modification of the SII, referred to as the Extended SII (ESII), has been shown to lead to much better performance in conditions with fluctuating interferers [124, 125]. The approach consists of evaluating the SII in short time-frames ranging from 9.4ms to 35ms, and averaging the resulting values [124]. The ESII has been shown to account accurately for the intelligibility of speech in sinusoidally amplitude modulated and interrupted noise, with modulation rates ranging from 4 to 128 Hz [125]. Several

other SIP algorithms have made use of short windows, with the same purpose in mind [7, 79, 140]. A different approach, presented in [25], hypothesizes that intelligibility is related to the presence of spectro-temporal regions with high SNR, or "glimpses". Is is shown that the number of such glimpses is indeed a good predictor of intelligibility in conditions where speech is masked by different numbers of interfering talkers [25].

Type 2M SIP Algorithms

The methods, discussed above, require clean speech and noise in separation, and therefore do not allow for predicting the intelligibility of non-linearly degraded speech. To allow for doing so, the type 2M algorithms base their predictions on 1) the degraded (i.e noisy and/or non-linearly processed) speech signal for which it is desired to predict intelligibility, as well as 2) a corresponding clean speech signal which is not affected by noise or other degradation.

An early SIP algorithm for predicting the intelligibility of non-linearly processed speech is the Speech Transmission Index (STI) [67, 68, 138], first introduced in 1971. The procedure assumes speech to be transmitted across a well-defined communication channel. The impact of this channel, on the modulations contained in the transmitted signal, is determined by transmitting a sequence of artificial probe signals across the channel, and monitoring the output [67]. The resulting measurements are used to compute an index, summarizing how well the channel conserves the modulations of transmitted signals [67]. Because the STI requires transmission of artificial probe signals, it does not fit directly into the taxonomy of Fig. 10. Furthermore, the use of artificial probe signals requires access to the communication channel, and therefore does not allow for predicting the intelligibility of pre-recorded degraded signals. This has led to the proposal of several STI-like methods which use speech as a probe signal [58]. These are type 2M algorithms (Fig. 10), as they make use of a clean speech signal (the probe) and a degraded speech signal (the probe, transmitted across the channel). The STI has been widely used, especially within architectural acoustics [143].

Since the introduction of the STI, several approaches for predicting the intelligibility of non-linearly degraded speech have been investigated. One such approach is to employ existing algorithms for predicting speech *quality* [120], assuming some degree of correlation between intelligibility and perceived quality. These algorithms have, however, shown limited correlation with speech intelligibility [101, 141]. Therefore, several type 2M algorithms, specifically for predicting intelligibility, have been proposed. The Spectro-Temporal Modulation Index (STMI), proposed in [41], takes inspiration from the STI, but uses a more complex auditory model, and considers modulations jointly across time and frequency. Another notable example is

the Coherence SII (CSII), which attempts to predict the impact of both noise and non-linear distortion that can occur in a hearing aid [83]. The CSII computes the SII, but uses the coherence [19] between clean and degraded signals instead of the SNR [83]. The more recent Hearing-Aid Speech Perception Index (HASPI), which can be considered an updated version of the CSII, includes a considerably more advanced auditory model, bases predictions on a larger set of features, and can account for individual hearing loss [84].

Another recent type 2M algorithm is the Short-Time Objective Intelligibility (STOI) measure [140], which has gained considerable popularity in the speech processing community due to its simplicity and high prediction performance. The STOI measure is computed as the sample correlation coefficient between short segments of clean speech envelopes and degraded speech envelopes, averaged across both 15 one-third octave bands and across time [140]. Despite the simplicity of this approach, the STOI measure has been shown to accurately account for the intelligibility of speech in conditions including different noise sources [140], noise reduction processing [134, 140], and noisy speech transmitted via telephone [77], as well as speech processed by hearing-aids and cochlear implants [46]. A number of alternate versions of the STOI measure have been proposed with the aim of improving prediction performance in different conditions [76, 99, 135, 136]. These include, for example, the Weighted STOI (WSTOI) measure, which uses information theory to weight each time-frequency region according to an estimate of its information content [99], and the Extended STOI (ESTOI) measure, which introduces additional computational steps to improve prediction performance in conditions with fluctuating interferers [76]. A speech enhancement algorithm, which maximizes the STOI measure, is proposed in [98].

Another group of popular SIP algorithms are based on the Envelope Power Spectrum Model (EPSM) [43], which accounts for observed thresholds in modulation detection tasks. The speech-based EPSM (sEPSM) [78] uses the EPSM to predict the intelligibility of speech, arguing that modulations are a key feature of speech. The sEPSM does not fit well into the taxonomy of Fig. 10, as it requires a degraded signal and a clean *noise* signal (as opposed to a clean speech signal). However, it is perhaps most closely related to the type 2M predictors. A short-time version, the multi-resolution sEPSM (mr-sEPSM), which better accounts for the intelligibility of fluctuating maskers is proposed in [79]. A number of variations of the sEPSM, with different specialized properties, are proposed in [20, 44].

2.2 Intrusive Prediction of Binaural Intelligibility

The SIP algorithms discussed in the previous section all assume that speech is presented either to one ear only, or that identical signals are presented to both ears. However, in most real-world scenarios, due to the environmental

acoustics and the acoustics of the head of the listener, different signals are received by the two ears. These differences can be separated in two components: Interaural Level Differences (ILDs) are differences in sound pressure from one ear to the other, and are responsible for the better-ear advantage discussed in Section 1.2, while Interaural Time Difference (ITD) are differences in arrival time from one ear to another, and are responsible for binaural unmasking. Collectively, ILDs and ITDs are known as interaural cues. As an example, if a point sound source is placed in front of a listener in an anechoic environment, the sound arrives simultaneously at both ears, and with equal intensity. Conversely, if the source is placed to the right of the lister, the sound arrives at the right ear before arriving at the left ear, and the signal at the left ear is less intense, because it is attenuated by the acoustics of the head (as illustrated in Fig. 9). This clearly shows that the interaural cues contain information about the location of the sound source, relative to the listener. As discussed in Section 1.2, when a speech signal is masked by a noise interferer, and the two are presented with different interaural cues (corresponding to different source positions), intelligibility is generally higher than for the diotic situation, due to binaural advantage [15, 17, 18, 33, 36, 113, 118]. In some conditions the binaural advantage can lead to SRTs which are 12-14 dB lower, compared with diotic presentation [18].

Modeling Binaural Advantage

Early efforts to model binaural processing in the brain were made for purposes unrelated to speech intelligibility. In 1948, Jeffress proposed an influential model of human sound localization [74], suggesting that the auditory system uses an array of delay lines to estimate the ITD, which is, in turn, used to estimate the direction of arrival. In 1963, Durlach proposed another famous model, which uses a similar mechanism to explain binaural advantage [38, 39]. As with the model by Jeffress, this model was not developed with speech intelligibility in mind, but rather to account for empirically observed binaural advantages in psychophysical experiments concerning the detection of tones in noise. The model suggests that the brain uses a so-called Equalization Cancellation (EC) stage to combine the signals from the left and right ears. The EC stage operates in two steps: 1) *equalization:* the left and right input signals are time-shifted and scaled relative to one-another, and 2) *cancellation:* the difference between the resulting signals is computed [38, 39]. The difference signal, computed in the cancellation step, is assumed to be used for all further auditory processing. Crucially, the applied time-shift and scaling parameters are determined such as to make the interferer signals identical in both ears. Thereby, the interferer signal becomes strongly attenuated in the cancellation step. This principle corresponds, essentially, to assuming that the auditory system is capable of

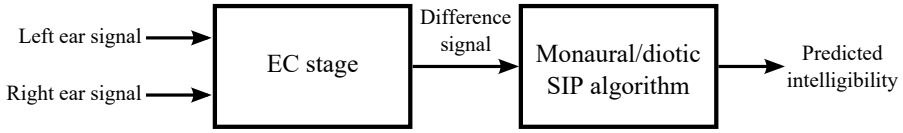


Fig. 11: Many binaural SIP algorithms are made by combining a monaural/diotic SIP algorithm with an EC stage. Intrusive SIP algorithms (type 1 and 2) require multiple input signals (clean speech and either noise or degraded speech). This requirement is fulfilled by using two parallel EC stages (not shown in the figure).

performing delay-and-subtract beamforming. In order to limit the predicted binaural advantage to human levels, random errors ("jitter") are added to the time-shifting and scaling parameters [38]. The EC stage can predict the outcome of a large range of psychophysical experiments relating to the detection of tones in noise [39].

A number of other models of binaural advantage have been suggested since the proposal of the EC stage [11–13, 24, 97, 155]. However, the EC stage remains highly influential and widely used today.

Type 1B SIP Algorithms

An early proposal of a binaural SIP algorithm was given in 1984 [145]. This was obtained by modifying the EC stage and combining it with the AI, as illustrated in Fig. 11. More recently, the method was revived and refined [6, 7]. The refined method, which combines the modified EC stage [145] with the more modern SII, is referred to as the Binaural Speech Intelligibility Measure (BSIM) [7]. The BSIM is able to account for the impact of individual hearing loss by adding audiogram-shaped stationary noise to the input signals [7]. A method for making predictions in conditions with fluctuating interferers, the short-time BSIM (stBSIM), is also proposed in [7]. This is based on the same principle as that underlying the ESII. The BSIM and stBSIM were shown to account reasonably well for normal-hearing SRTs in SSN (BSIM/stBSIM), babble noise (stBSIM), and for a single interfering talker (stBSIM), in four rooms with different reverberation times [7]. Furthermore, the methods were able to account for approximately 70% of the variance observed in the SRTs of hearing impaired listers relative to the SRTs of normal hearing listeners [7]. A number of extensions to the BSIM have later been proposed with the aim of improving predictions in reverberant environments [122, 123].

Several other binaural SIP algorithms have been proposed. These generally follow the same approach of combining an EC stage with a monaural/diotic SIP algorithm (see Fig. 11). An algorithm which is very similar to the BSIM is proposed in [150]. This differs by drawing specific realizations of jitter for each sample of the input signals [150], while the BSIM [7]

assumes the jitter to be constant across the entire length of the signal, and computes the expectation across this jitter instead of drawing realizations. A short-time version of the algorithm from [150] is introduced in [151], relying on a principle similar to the ESII.

A different branch of binaural SIP research focuses on a closed form expression for binaural advantage, which is derived from the EC stage [28, 29]. A binaural SIP algorithm, based on this expression, is proposed in [93]. The algorithm predicts binaural advantage by 1) computing the SII in a better-ear fashion, by using the ear with the maximum SNR, separately in each frequency band, and 2) adding binaural advantage according to the expression given in [28, 29]. A number of variations of this model are proposed in [75, 94, 95].

Type 2B SIP Algorithms

The binaural SIP algorithms, discussed above, all belong to the type 1B algorithms (Fig. 10), because they require speech and noise signals in separation. Therefore, they are not able to make predictions for non-linearly processed signals. A number of type 2B methods have therefore been proposed; all based on extending type 2M algorithms with an EC stage (i.e. following Fig. 11). These can, in principle, account for both non-linear processing and binaural advantage. A binaural version of the STI is proposed in [144]. A binaural version of the mr-sEPSM, the Binaural sEPSM (B-sEPSM), is proposed in [20]. The B-sEPSM predicts binaural intelligibility from degraded speech and clean *noise*, as recorded at the left and right ears of the listener. Similar to the sEPSM, the B-sEPSM does not technically fit directly into the taxonomy given by Fig. 10. However, since it does allow for predicting intelligibility for non-linearly processed signals, it is most closely related to the type 2B category. The B-sEPSM was shown to predict intelligibility well in binaural environments with one or more sources of SSN, speech modulated SSN, babble, or reversed speech [20]. The work in [20] does not evaluate the performance of the B-sEPSM for conditions with non-linear processing. However, the mr-sEPSM has previously been demonstrated to work well in conditions with spectral subtraction [79], while the sEPSM has been shown to predict the impact of transmitting noisy speech via telephone [77]. Due to the similarity of the methods in the sEPSM-family, it is likely that the B-sEPSM would be able to handle conditions with binaural advantage and non-linear processing.

2.3 Non-Intrusive Intelligibility Prediction

In some conditions, the clean signal, required for both type 1 and 2 predictors, cannot be obtained. This could be because the degraded signal is not based on a well defined clean speech signal (an example could be, to predict

the intelligibility of poorly spoken or synthesized speech). In other cases, the degraded signal could be distorted in some way that makes it difficult to compare it directly with the clean signal. For instance, many SIP algorithms function by comparing short adjacent segments of clean and degraded speech in narrow frequency bands (e.g. the ESII, the STOI measure, and the mr-sEPSM). This requires the clean and degraded signals to be strictly aligned in both time and frequency. If the degraded speech signal has been warped or translated in time or frequency (e.g. by changing the playback rate), it may no longer be possible to properly align the clean and degraded signals. Most of the algorithms discussed until this point would yield inaccurate predictions in this scenario. We are also likely to find ourselves in conditions where a clean signal could be reasonably defined, but simply is not available. For instance, we may wish to predict the intelligibility of previously recorded noisy speech. In such a case, a clean speech reference is typically not available. Furthermore, it could be desirable to run SIP algorithms as part of a real-time audio processing system. For instance, a hearing aid could use a SIP algorithm to continuously adapt its processing. We investigate this use-case in Section 3.2.

The above examples illustrate the need for *non-intrusive* SIP algorithms, i.e. SIP algorithms which do not require knowledge of a clean speech signal at all, but predict intelligibility based on the degraded signal alone. Fig. 10 refers to these as type 3 predictors. While the field of non-intrusive SIP is considerably less studied than the intrusive counterpart, several methods have been proposed. These operate according to a number of different principles.

Type 3M SIP Algorithms

We have identified five overall approaches, which have been investigated, for developing non-intrusive SIP algorithms.

One approach is to use one of several existing non-intrusive *quality* prediction algorithms, such as the ITU-T P.563 [3] or the ANIQUE+ [4, 85]. These algorithms extract and combine a large number of signal characteristics which have shown to correlate well with perceived audio quality. However, speech intelligibility and speech quality are not as similar as one might suspect, and the two aforementioned quality predictors have been shown to correlate poorly with speech intelligibility [46, 47].

A second approach to non-intrusive SIP uses the Speech to Reverberation Modulation energy Ratio (SRMR), proposed in [47]. The SRMR attempts to predict both quality and intelligibility of reverberant speech, based on the observation that clean speech predominantly contains energy at low modulation frequencies, while reverberated speech contains relatively more energy at higher modulation frequencies [47] (see also [42]). The SRMR is computed as a ratio of low-frequency modulation energy to high-frequency modula-

tion energy and has been shown to correlate well with the intelligibility of reverberant and de-reverberated speech [47]. Special versions of the SRMR have been developed for hearing impaired listeners [45] and cochlear implant users [45, 126]. Additional refinement and validation of the SRMR is given in [46, 127]. The average modulation-spectrum area (ModA) measure, proposed in [21], operates according to a similar principle. The ModA measure assumes intelligibility to be correlated with the area under the modulation spectrum [21]. This has been shown to correlate well with intelligibility in noisy and reverberant conditions [21, 46].

A third approach to non-intrusive SIP is to use an Automatic Speech Recognition (ASR) system to transcribe the degraded speech, and somehow measure the accuracy of the transcription. This is arguably not a truly non-intrusive approach, as it requires a transcription of the speech material. Such an approach is studied in [129]. By limiting the speech corpus to a single matrix sentence test, a Hidden Markov Model (HMM) based ASR system obtains performance which is remarkably similar to human performance. This method was shown able to predict the difference in intelligibility of matrix sentence tests in different languages [129] (this difference is observed in [63]). However, in its current format, the method is limited to predicting intelligibility for matrix sentence tests.

A fourth approach is to estimate the clean signal and use this to evaluate an intrusive SIP algorithm. A method for real-time non-intrusive SIP in hearing-aids is proposed in [135]. This method uses a multi-microphone beamformer, aimed at the target speaker, to estimate the clean signal [135]. The resulting signal clean signal is used, together with the degraded one, to evaluate the STOI measure. A refined version of the method, using also pitch-features for the clean speech estimation, is proposed in [136].

A fifth approach is to non-intrusively predict the output of an intrusive SIP algorithm, using machine learning. Predictions are typically based on a considerable number of different features extracted from the degraded speech signal. Machine learning approaches investigated for this purpose include Gaussian mixture models [131], hidden Markov models [80], and tree based regression [132]. It has generally been found to be possible to predict the output of various intrusive measures accurately without using the clean signal. A disadvantage of this approach is that prediction performance is inevitably limited by the performance of the underlying intrusive algorithm.

Type 3B SIP Algorithms

We are only aware of one effort towards non-intrusive binaural (type 3B) SIP, obtained by extending the SRMR with an EC stage [27]. This was found to account well for the intelligibility of speech masked by a single point source of stationary noise, at different positions, in anechoic and reverber-

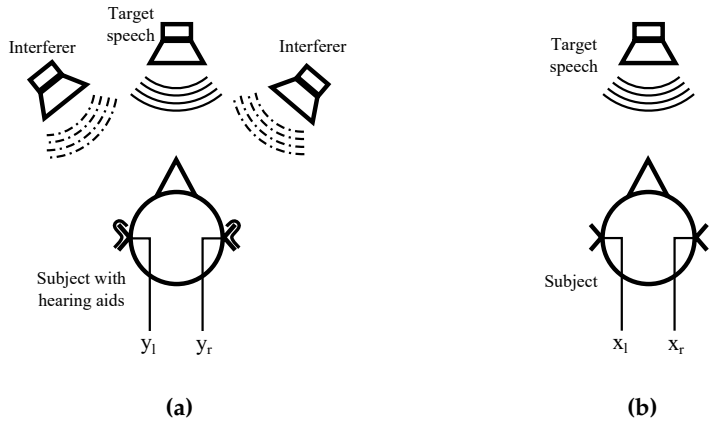


Fig. 12: An illustration of how one could define binaural degraded and clean signals in relation to hearing aid processing. **(a)** The degraded signals are recorded in the ear canals of the subject in the presence of target speech and interferers, following hearing aid processing. **(b)** The clean signals are also recorded in the ear canals of the subjects, but without interferers and without hearing aid processing. Figure adapted from paper [D].

ant rooms [27].

3 Applications to Hearing Aid Systems

In the work underlying this thesis, we have studied SIP with the aim of eventual applications within the development and operation of hearing aid systems. In this section we discuss two possible use-cases of SIP within this field. Before doing so, we consider the signals involved when applying SIP for hearing aid systems.

A typical difficult listening situation for a hearing aid user is illustrated in Fig. 12a. Here, the user is wearing a pair of hearing aids, and attempts to understand speech from a talker, while being present in an environment that contains one or more noise sources, and possibly reverberation. In such a case, sound from several different sources impinge on the head of the listener from different directions. The combination of speech, noise, and reverberation is impacted by the acoustics of the human head, and is picked up by one or more hearing aid microphones, located in or close to each ear of the listener. The hearing aids then carry out various forms of processing, which may include beamforming, noise reduction, speech enhancement, feedback cancellation, amplitude compression, and amplification [32, 81, 130]. Finally, the hearing aid uses a small loudspeaker to present the processed signal to the user.

In the present context, we are interested in predicting the intelligibility of

the output of a hearing aid, as measured in the ear canal of the wearer. As already remarked, there are several non-linear processing stages in the hearing aid. It is not, in general, possible to separate speech and noise components after this type of processing has been applied. Therefore, the intelligibility of a hearing aid processed signal cannot be predicted by a type 1 SIP algorithm, as this requires knowledge of speech and noise in separation. It may be possible to use a type 2 algorithm, which requires degraded speech (i.e. the signal in the ear canal of the listener, including speech, noise and hearing aid processing; see Fig. 12a) and clean speech, provided that it is possible to obtain a clean speech reference signal. In this thesis, we have defined the clean speech signals to be those recorded in the ear of the listener, in the absence of noise and hearing aid processing, but in the presence of speech (as illustrated in Fig. 12b). Such a signal is not available to a hearing aid in typical usage scenarios. However, it may be possible to record or simulate the clean signal in some controlled experiments. In cases where the clean signal is not available, a type 3 algorithm may be used to predict intelligibility from the degraded signal alone.

In the following, we consider two use-cases of SIP algorithms within hearing aid development.

3.1 Predicting the Outcome of Listening Experiments

The first use-case we consider is to predict the outcome of listening experiments within the process of developing new hearing aids.

The General Use-Case

When hearing aid manufacturers develop new generations of hearing aids, they strive to make them perform better than previous generations. Performance of hearing aids can be quantified in numerous ways such as sound quality, comfort, looks or visibility, battery life, etc. It is obvious that speech intelligibility is an extremely important performance measure, as improved intelligibility is, perhaps, the primary reason for wearing hearing aids. Therefore, many listening experiments are typically carried during the course of developing a hearing aid. In these experiments, SRTs are measured for a number of hearing impaired test subjects wearing different hearing aids. The set of tested hearing aids typically include novel prototypes as well as hearing aids from the previous generation, representing the "state-of-the-art". Each of these may be tested in multiple settings, i.e. for different levels of noise reduction, directionality, etc.

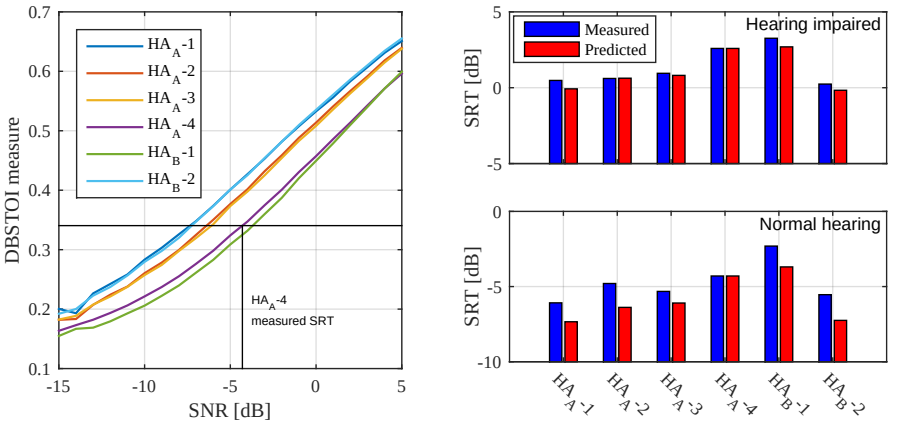
Listening experiments, carried out as described in Section 1.2, are a powerful means to gain knowledge about the extent to which intelligibility is improved by newly developed prototypes. However, listening experiments

are also time consuming and expensive to carry out: they require weeks or months of preparation, expensive equipment, recruitment of test subjects, and careful data analysis. These troubles may be entirely worthwhile, if important learnings are extracted from the experiment. However, at the same time, it is obviously desirable to keep the number of listening experiments to the bare minimum. One way to do so, could be to replace a subset of listening experiments with predictions made by SIP algorithms. We suggest two different approaches for doing so:

1. Make predictions based on *actual* audio recordings from the environment in which the experiment would have taken place. These recordings can be made using a Head And Torso Simulator (HATS), which has microphones mounted in the ear canals. Segments of degraded speech can then be recorded for all the hearing aids and hearing aid settings included in the experiment, in a range of SNRs.

2. Make predictions based on *simulated* signals. These can be made using computational models of both the hearing aids, and the environment in which the test would have taken place. The environment may be simulated with a room acoustical model or with prerecorded Room Impulse Responses (RIRs). A software implementation of the hearing aid processing is used for simulating the hearing aid.

The first option ensures signals which are highly consistent with those which would have been experienced by a subject in an actual listening experiment, while the second approach is dependent on the accuracy with which the hearing aids and the acoustical environment can be simulated. Conversely, the first method requires an experimenter to carry out many of the preparations needed for a listening experiment. In particular, it is required that a sound studio (or another environment) is fitted with the necessary equipment for playback and recording and, furthermore, that the tested hearing aids are available and correctly programmed. The second approach is entirely software-based, and it could potentially run in an automated fashion, e.g. on a nightly basis, and thus keep developers constantly informed about the current predicted intelligibility performance of the developed prototypes. Both means of obtaining signals are reproducible: it is possible to first record degraded signals, including noise and hearing aid processing, and then record clean signals for the same speech material, without noise and hearing aid processing. This allows the use of type 2 algorithms for making speech intelligibility predictions. This is a considerable advantage, as type 3 (non-intrusive) algorithms should be expected to yield much less accurate results, because they have much less information available to base predictions on.



(a) **(b)**

Fig. 13: (a) The DBSTOI measure plotted against input SNR, for signals measured in the ears of a HATS, wearing two different pairs of hearing aids in a total of six different settings. The average SRT measured for normal hearing listeners, wearing $\mathrm{HA_A}$-4, is shown as a vertical black line. A horizontal line shows the corresponding DBSTOI measure (which, in this case, equals approximately 0.34). **(b)** Measured and predicted SRTs for hearing impaired and normal hearing listeners. Predictions were made by calibrating to the measured SRT for $\mathrm{HA_A}$-4. The SRT predictions for normal hearing listeners can be read off from (a) as the SNRs, at which the curves intersect the horizontal black line.

A Specific Example

To further substantiate the proposal of replacing listening experiments with predictions, we give a brief account of an otherwise unpublished project, in which we attempted to predict the outcome of an experiment, carried out by colleagues at Oticon A/S as part of the development of a new product. In this experiment, SRTs were measured for 14 hearing impaired listeners, wearing two different hearing aids in a total of six different settings. The subjects were experienced hearing aid users with approximately symmetrical hearing losses ranging from mild to severe. The hearing aids were individually fitted, such as to compensate for the hearing loss of each subject, and were otherwise configured exactly as they would have been for normal use, including the use of non-linear processing steps such as amplitude compression and feedback cancellation. The tested settings spanned different combinations of noise reduction and directionality. The specific hearing aids and their settings are, unfortunately, confidential, and are therefore simply referred to as $\mathrm{HA_A}$-1, $\mathrm{HA_A}$-2, $\mathrm{HA_A}$-3, $\mathrm{HA_A}$-4, $\mathrm{HA_B}$-1, and $\mathrm{HA_B}$-2 (i.e. $\mathrm{HA_A}$ was tested in four settings and $\mathrm{HA_B}$ was tested in two settings). The experiment was carried out in a sound studio with acoustically treated walls, using the Danish Dantale II matrix sentences [149]. Speech was presented from the front (i.e. 0° azimuth),

and was masked by two interferers, placed at 15° and 180° azimuth (i.e. one slightly to the right of the front, and one directly behind the listener). The interferers consisted of different segments of the International Speech Test Signal (ISTS); an artificial speech-like but unintelligible signal, composed of short segments of speech in different languages [65].

Since the SIP algorithms investigated in this thesis do not include the ability to explicitly model individual hearing losses (e.g. based on individual audiograms), we found it overly ambitious to believe that accurate predictions could be made for hearing impaired listeners. Therefore, we chose to repeat the actual listening experiment, using 14 normal hearing listeners. These were fitted with hearing aids providing only minor amplification (they were programmed to compensate for a hearing loss of 20 dB HL at low frequencies, and 40 dB HL at high frequencies, transitioning linearly from 500 Hz to 2 kHz). By doing so, we were able to separately asses prediction performance for normal hearing and hearing impaired subjects.

In order to allow for predicting intelligibility, we made recordings using a Brüel & Kjær Head And Torso Simulator (HATS), wearing hearing aids, and placed where the subjects had been seated during the experiment. We recorded a fixed sequence of 30 Dantale sentences for 21 SNRs, evenly spaced by 1 dB from -15 to +5 dB. This sequence of recordings was repeated for all six hearing aid settings. Furthermore, one clean recording of the sequence, excluding noise and hearing aids, was made. Thus the 30 sentences were recorded a total of $21 \times 6 + 1$ times.

To predict intelligibility, we computed the Deterministic BSTOI (DBSTOI) measure for each of the degraded recordings. The DBSTOI measure is a type 2B algorithm developed as part of the work underlying this thesis (see Section 4 or papers [B] and [D]). This resulted in "scorings" of intelligibility on an arbitrary, unit-less, scale from 0 to 1, as shown in Fig. 13a. This scale has empirically been shown to have a nearly monotonic correspondence with measured intelligibility across a large range of conditions (assuming a fixed speech corpora and a fixed group of listeners). By using this property, it is possible to predict the SRT in one condition from knowledge of the SRT in another one. We, arbitrarily, picked $\mathrm{HA}_{\mathbf{A}}$-4 as a reference condition, and attempted to predict the SRTs of the other conditions, as illustrated in Fig. 13a:

1. **Calibration:** The DBSTOI measure was evaluated at the SRT of $\mathrm{HA}_{\mathbf{A}}$-4, as shown by the vertical black line in Fig. 13a. The exact value was obtained by linear interpolation between adjacent points. Assuming a monotonic correspondence between speech intelligibility and the DBSTOI measure, this value of the DBSTOI measure corresponds to the SRT, in any condition. The resulting calibration value is marked by a horizontal line in Fig. 13a.

2. **Prediction:** The SRTs in other conditions were estimated, by finding

the SNRs, at which these conditions give rise to a DBSTOI measure equal to the calibration value. In Fig. 13a, this is seen as the intersections between the lines representing values of the DBSTOI measure, and the horizontal black line representing the calibration value.

The resulting predicted SRTs for normal hearing (as obtained in the repeated experiment) and hearing impaired listeners (as obtained in the original experiment) are shown in Fig. 13b. Notice that different predictions are made for normal hearing and hearing impaired listeners. This is because the calibration step accounts for the differences in average performance between the two groups (in a reference condition). Somewhat surprisingly, accurate predictions are made for both normal hearing and hearing impaired listeners, in spite of the fact that only a single calibration value is used to account for the difference in hearing fidelity between the two groups. In fact, predictions for hearing impaired listeners are more accurate than predictions for normal hearing listeners. Possibly, the larger prediction errors for normal hearing listeners could be caused by random measurement errors: the used procedure for estimating SRTs is particularly sensitive to errors in the condition used for calibration. Since the SRTs are overestimated for every condition with normal hearing listeners, it is likely that prediction performance could have been considerably improved by picking a different condition for calibration. It is worth noting that, regardless of the absolute prediction errors, the DBSTOI measure is able to accurately determine which hearing aids and settings give rise to relatively better intelligibility, and which give rise to relatively lower intelligibility.

The above results suggest that SIP algorithms could be highly useful within hearing aid development. In particular, it was possible to accurately predict the outcome of a listening experiment with hearing impaired test subjects, carried out as part of the product development within Oticon A/S. This suggests the possibility of partially replacing time consuming and expensive listening experiments with much cheaper predictions. The predictions shown here were based on studio recordings. While these are neither quick or free to make, the involved expense is far below that of a conventional listening experiment. It is furthermore likely, that it will eventually be possible to use entirely simulated signals, in place of studio recordings.

3.2 Hearing Aid Systems that Predict Intelligibility

A somewhat more ambitious application of SIP, is to predict intelligibility on a hearing aid system in real-time during usage, and use the output to optimize the operation of the system. Similar applications have previously been discussed in [46, 135, 136].

A possible strategy for using a SIP algorithm to guide the operation of a hearing aid system, could be to trade between quality and intelligibility.

In difficult situations with high noise levels, the hearing aid user will most likely want the device to apply as much noise reduction and directionality as possible, to ensure a certain minimum of intelligibility. However, in easier, quieter conditions, the same noise reduction and directionality systems may lead to unnatural sound and a loss of directional cues [31, 87]. A hearing aid system including a SIP algorithm, could be designed to use these systems only when they are necessary to obtain a predefined minimum of intelligibility. Thereby, the user could experience natural sound in quieter conditions, while being aided by increasingly aggressive processing in more difficult situations.

There are two obstacles which make this use-case more ambitious than the one discussed above:

1. When a SIP algorithm runs in real-time on a hearing aid system, there is no obvious way to obtain a clean signal. Thus, only non-intrusive (type 3) SIP algorithms are applicable for this use.

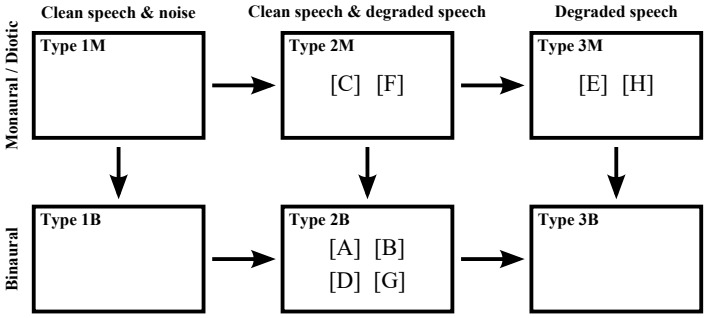
2. The amount of computational resources available to a hearing aid system is typically limited by severe constraints on size and power consumption. A non-intrusive SIP algorithm, intended for use in a hearing aid system, therefore ideally has to be computationally cheap to evaluate.

Given the current level of research within non-intrusive SIP, we expect it to be possible to fulfill both requirements in the near future.

4 Summary of Contributions

The main body of this thesis (Part II) consists of a collection of eight papers. The work underlying these papers can collectively be viewed as an effort to develop SIP algorithms which are suitable for the two applications discussed in Section 3. In order to reach this goal, we used the STOI measure [140] (monaural/diotic and intrusive, type 2M) as a starting point, because of its simplicity and proven record of high prediction accuracy across a broad range of conditions.

To facilitate the applications suggested in Section 3, we extended the STOI measure along mainly two lines: 1) binaural intelligibility prediction (papers [A], [B], [D] and [G]), and 2) non-intrusive intelligibility prediction (papers [E] and [H]). Furthermore, two smaller studies investigate diotic and intrusive extensions to the STOI measure (papers [C] and [F]). Fig. 14 places the work in relation to the taxonomy of SIP algorithms proposed in Fig. 10.



[A] "A binaural short time objective intelligibility measure for noisy and enhanced speech," in *Proc. Interspeech*, 2015.

[B] "A method for predicting the intelligibility of noisy and non-linearly enhanced binaural speech," in *Proc. ICASSP*, 2016.

[C] "Speech intelligibility prediction as a classification problem," in *Proc. MLSP*, 2016.

[D] "Predicting the intelligibility of noisy and nonlinearly processed binaural speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[E] "A non-intrusive short-time objective intelligibility measure," in *Proc. ICASSP*, 2017.

[F] "On the use of band importance weighting in the short-time objective intelligibility measure," in *Proc. Interspeech*, 2017.

[G] "Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions," *submitted to Speech Communication*, 2017.

[H] "Non-intrusive speech intelligibility prediction using convolutional neural networks," *submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

Fig. 14: The eight papers, which constitute the main body of this thesis, propose different SIP algorithms. The figure shows the papers plotted into the taxonomy of Fig. 10, according to the properties of the proposed methods.

4.1 Specific Contributions

In the following, we shortly summarize the purpose, methods and findings of each paper.

[A] A Binaural Short Time Objective Intelligibility Measure for Noisy and Enhanced Speech

In this paper we propose a binaural version of the STOI measure: the Binaural STOI (BSTOI) measure. This is achieved by extending the STOI measure with an EC stage in a manner similar to that proposed in [6, 145]. The EC stage combines left and right ear signals by relatively time shifting and scaling the signals, and thereafter subtracting them from one-another (as described in Section 2.2). In doing so, it introduces random time and amplitude jitter, which serves to limit the predicted binaural advantage to levels which are consistent with human performance. In the BSTOI measure, realizations of

this jitter are drawn, and the STOI measure is computed for each realization and averaged (i.e. the jitter is accounted for by use of Monte Carlo simulation).

Using measurements from three different listening experiments, we show that the BSTOI measure can accurately predict 1) binaural intelligibility in anechoic conditions with a frontal talker and a single SSN interferer placed in the horizontal plane, 2) intelligibility of diotic noisy speech processed with Ideal Binary Mask (IBM) proecssing, and 3) binaural intelligibility in conditions with multiple interferers and beamforming. Notably, the results suggest that the diotic prediction performance is not significantly decreased by the addition of the EC stage. Furthermore, the results show that accurate predictions are made in both simple conditions with large binaural advantage, and in slightly more acoustically complex conditions.

[B] A Method for Predicting the Intelligibility of Noisy and Non-Linearly Enhanced Binaural Speech

A disadvantage of the BSTOI measure is the use of Monte Carlo simulation for handling the jitter introduced by the EC stage. Because of this, the STOI measure has to be evaluated many times for different realizations of the jitter. This causes the BSTOI measure to be computationally expensive to evaluate, and thereby severely limits its usefulness in a range of applications. Furthermore, in spite of averaging across the realizations of jitter, the output remains somewhat noisy (i.e. slightly different outputs are obtained when computing the measure multiple times for the same input signals). In this paper we propose an improved version of the BSTOI measure, which does not rely on Monte Carlo simulation: the Deterministic BSTOI (DBSTOI) measure. To do so, we introduce two modifications of the original STOI measure: 1) the removing of a "clipping stage", and 2) the use of signal envelopes squared instead of conventional signal envelopes. This makes it possible to derive a closed-form expression for the expectation of the STOI measure across the jitter, thus removing the necessity of Monte Carlo simulation.

The DBSTOI measure is shown to make predictions in less than one tenth of the computational time required by the BSTOI measure. Additionally, it is demonstrated that the DBSTOI measure performs similarly to the BSTOI in terms of predicting the binaural intelligibility 1) in anechoic conditions with a frontal talker and a single SSN interferer placed in the horizontal plane, and 2) in conditions with multiple interferers and beamforming. Furthermore, an additional listening experiment was carried out as part of the work. This experiment measured the intelligibility of frontal speech masked by a point source of either SSN or "bottling factory hall noise". Intelligibility was measured for the unprocessed stimuli, and when applying IBM processing separately on the left and right ears of the listener. The DBSTOI measure

is shown to accurately predict the results of this experiment, including the conditions which involve both binaural advantage and non-linear processing (i.e. IBM processing) combined.

[C] Speech Intelligibility Prediction as a Classification Problem

The STOI measure predicts average intelligibility from at least several seconds of speech. However, much can be learned about speech, from the factors which make single words or phonemes intelligible or unintelligible. Paper [C] therefore works towards a version of the STOI measure which can predict intelligibility for such individual units of speech. This is referred to as a *microscopic* SIP algorithm [25]. The proposed method uses STOI-like features, together with linear discriminant analysis, to predict the intelligibility of individual *logatomes*, short tokens of non-sense speech [103], presented together with competing speech. The method can, to a considerable extent, account for the intelligibility of individual logatome tokens.

[D] Predicting the Intelligibility of Noisy and Non-linearly Processed Binaural Speech

In this paper, we further analyze and investigate the DBSTOI measure, proposed in paper [B]. Firstly, the expectation of the STOI measure across the jitter is explicitly derived (this derivation was left out in paper [B]). Secondly, the mathematical properties of the DBSTOI measure are more closely investigated. Specifically, it is shown that the DBSTOI measure reduces to the STOI measure for diotic stimuli, i.e. the EC stage plays no role when the inputs to the left and right ears are identical. This asserts that the addition of an EC stage is not detrimental to the performance of the underlying SIP algorithm. Thirdly, it is shown that the the DBSTOI measure performs similarly to the STOI measure with respect to predicting the intelligibility of diotic noisy speech processed with IBM processing. Since the EC stage becomes unimportant in diotic conditions, this result only suggests that the modifications introduced to the STOI measure, in order to allow for deriving the closed-form expectation across the jitter, are not detrimental to performance.

[E] A Non-Intrusive Short-Time Objective Intelligibility Measure

In this paper, we propose a non-intrusive version of the STOI measure: the Non-Intrusive STOI (NI-STOI) measure. This is done by estimating the clean speech envelopes from the degraded ones, thus eliminating the necessity of a clean reference signal. The clean signal envelopes are estimated by projecting the modulation magnitude spectrum of the degraded signal into a subspace spanning only modulation magnitude spectra which are present in speech. This subspace is generated in advance, using a database of clean speech.

The NI-STOI measure is demonstrated to correlate well with the measured intelligibility of diotic noisy speech processed with IBM proecssing. Predictions made using the NI-STOI measure are shown to be considerably more accurate than predictions made with the (non-intrusive) SRMR measure, but, as expected, less accurate than predictions made with the (intrusive) STOI measure.

[F] On the use of Band Importance Weighting in the Short-Time Objective Intelligibility Measure

Most conventional SIP algorithms include a Band Importance Function (BIF), specifying the relative contribution of each frequency band to intelligibility. The STOI measure assumes a uniform importance of the 15 considered one-third octave bands. This paper investigates whether the STOI measure can be improved by adding a fitted BIF. BIFs are fitted such as to maximize prediction performance on a number of different sets of measured intelligibility. This results in a number of candidate BIFs. These are all cross-evaluated on the different sets of measured intelligibility. However, none of the found candidate BIFs show a convincing improvement in prediction performance compared with the uniform one. This result suggests that the uniform BIF, used in the original STOI measure, is a reasonable choice.

[G] Refinement and Validation of the Binaural Short Time Objective Intelligibility Measure for Spatially Diverse Conditions

Some of the results, presented in paper [D], suggest that the DBSTOI measure has a tendency to overestimate intelligibility in conditions with multiple spatially distributed interferers. Paper [G] therefore further investigates the performance of the DBSTOI measure in such environments. To do so, an additional dataset of measured intelligibility was collected, consisting of conditions with mixtures of point-source SSN and cylindrically isotropic SSN. Furthermore, another dataset, including conditions with both non-linear processing and multiple spatially distributed interferers, was obtained. Predictions for these datasets clearly show that the DBSTOI overestimates intelligibility considerable in conditions with spatially distributed interferers. A detailed investigation of this problem reveals this to be caused by an upwards bias, stemming from the combined usage of the EC stage and the STOI measure. Based on this insight, we derived a version of the binaural STOI which does not suffer from this issue: the Modified BSTOI (MBSTOI) measure. This is shown to compare favorably with both the DBSTOI measure and the B-sEPSM, across five datasets of measured intelligibility.

[H] Non-Intrusive Speech Intelligibility Prediction using Convolutional Neural Networks

In this paper, we investigate the use of Convolutional Neural Networks (CNNs) for non-intrusive SIP. The applied CNN architecture was designed specifically with a purpose of being interpretable and easy to visualize, and furthermore shows strong similarities to existing SIP algorithms.

To evaluate the performance of the proposed method, we collected four datasets of diotic intelligibility from the literature, recreated or adapted them for the purpose at hand, and combined them into one large dataset. Different approaches for splitting this dataset into training, validation, and testing subsets were investigated. Prediction performance was compared to two existing intrusive methods, the STOI and ESTOI measures, as well as to existing non-intrusive methods, the NI-STOI measure and the SRMR. In the investigated conditions, the proposed method performed similar to or better than the four existing methods. Lastly, a specific instance of a trained network is investigated in depth, by visualizing the fitted weights. This illustrates how the network predicts intelligibility by detecting the presence of a number of speech-like modulation structures.

4.2 Summary of Conclusions

The primary outcome of the work underlying this thesis, has been 1) the proposal of a binaural intrusive version of the STOI measure (type 2B), and 2) the proposal of a non-intrusive version of the STOI measure (type 3M).

Throughout the work, multiple refined versions of the binaural STOI measure have been proposed. The final version, the MBSTOI measure, was shown able to predict the impact of a large number of different factors, including different noise types, different spatial arrangements of interferers ranging from single point sources to isotropic noise, different types of linear and non-linear processing, and reverberation. Furthermore, a previously unpublished experiment, described in this introduction, showed the potential of the binaural STOI measure as an alternative to time consuming and expensive listening experiments.

A second topic was the development of a non-intrusive version of the STOI measure. This was approached from two different directions. In the first approach, a model of clean speech was used to estimate the required clean speech features from the noisy speech features. This made it possible to evaluate the intrusive STOI measure, without the need for a clean reference signal. In the second approach, STOI-like features were used together with a CNN to predict intelligibility. Both solutions have shown high prediction accuracy in comparison with other non-intrusive and intrusive intelligibility

measures. However, at the same time it is evident, that more work is required before non-intrusive intelligibility prediction can become useful in real-time speech processing systems.

In broader terms, this thesis has investigated the potential of SIP algorithms in the STOI family, for use within the development and operation of hearing aid systems. The results firmly suggest that some listening tests may be replaced by cheaper and less time consuming predictions. Further into the future, we find it likely that real-time implementations of non-intrusive SIP algorithms will find use in hearing aid systems. Our results underline that the STOI family of SIP algorithms is a good starting point for reaching both of these targets.

5 Directions of Future Research

The binaural version of the STOI measure, proposed in this work, has already been considerably matured, and we consider it to be ready for use in real-world applications. Specifically, the method has shown highly reliable within the domain spanned by the experimental results at our disposal. We therefore believe the most important next step to be for scientists and practitioners to gain practical experience with the use of the method. Such experience will certainly lead to a better understanding of 1) when predictions can be trusted to be accurate and when they cannot, and 2) what further modifications are necessary to make potential future revisions of the binaural STOI measure even more trustworthy.

Historically, the non-intrusive SIP problem has not been investigated nearly as much as the intrusive one. In this work we demonstrated that the performance of non-intrusive SIP algorithms can approach that of intrusive ones, under appropriate conditions. We did so, using both a traditional type of SIP algorithm, a non-intrusive version of the STOI measure, and using CNNs. However, much more evaluation should be carried out, to establish how broadly these algorithms can be expected to work. As larger datasets of measured intelligibility become available, approaches involving deep learning become increasingly interesting. Deep learning has had enormous success within e.g. automatic speech recognition, and we firmly believe that similar results can be obtained in using it for SIP.

The methods considered in this thesis include no steps to model individual hearing loss. We deliberately excluded this topic to limit the scope of the work. Somewhat surprisingly, it was shown that average intelligibility could be accurately predicted for a group of hearing impaired listeners, by adequately calibrating for the general performance of the group. However, it is unlikely that this extremely simple approach will be sufficient for all purposes. It therefore remains as an important piece of future work to develop

SIP algorithms which can accurately account for individual hearing loss.

Lastly, we express a note on openness. It is crucial that future SIP algorithms can be compared to previous ones, and that this can be done for the greatest possible number of listening conditions. We therefore urge fellow researchers to share implementations of proposed SIP algorithms to the furthest possible extent. Ideally this should be done publicly and freely. Furthermore, we suggest that future projects, concerning SIP, consider allocating resources to the collection of large and well-documented databases of measured speech intelligibility, which are shared openly and freely. While this is not always possible to do, it is perhaps one of the most effective ways of contributing to the future advancement of SIP.

References

[1] "Methods for the calculation of the articulation index," American National Standards Institute, New York, United States, Standard, 1969, ANSI S3.5:1969.

[2] "Methods for calculation of the speech intelligibility index," American National Standards Institute, New York, United States, Standard, 1997, ANSI S3.5:1997.

[3] "Single-ended method for objective speech quality assessment in narrow-band telephony applications," International Telecommunication Union, Standard, 2004, ITU P.563.

[4] "Auditory non-intrusive quality estimation plus (ANIQUE+): Perceptual model for non-intrusive estimation of narrowband speech quality," Alliance for Telecommunications Industry Solutions, Standard, 2006, ATIS 0100005.

[5] S. P. Bacon, J. M. Opie, and D. Y. Montoya, "The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds," *Journal of Speech, Language, and Hearing Research*, vol. 41, no. 3, pp. 549–563, Jun. 1998.

[6] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 331–342, Apr. 2006.

[7] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension and evaluation of a binaural speech intelligibility model," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2479–2497, Dec. 2010.

[8] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 1996, ISBN: 9780262024136.

[9] Blausen.com staff, "Medical gallery of blausen medical 2014," Retrieved from https://en.wikipedia.org/wiki/Respiratory_tract#/media/File:Blausen_0872_UpperRespiratorySystem.png on 08/07-2017.

[10] T. Brand and B. Kollmeier, "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," *J. Acoust. Soc. Am.*, vol. 111, no. 6, pp. 2801–2810, Jun. 2002.

[11] J. Breebaart and S. van de Par, "Binaural processing model based on contralateral inhibition. I. model structure," *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 1074–1088, Aug. 2001.

[12] ——, "Binaural processing model based on contralateral inhibition. II. dependence on spectral parameters," *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 1089–1104, Aug. 2001.

[13] ——, "Binaural processing model based on contralateral inhibition. III. dependence on temporal parameters," *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 1105–1117, Aug. 2001.

[14] A. W. Bronkhorst, "The cocktail-party problem revisited: early processing and selection of multi-talker speech," *Attention, perception & psychophysics*, vol. 77, no. 5, pp. 1465–1487, Jul. 2015.

[15] A. W. Bronkhorst and R. Plomp, "The effect of head-induced interaural time and level differences on speech intelligibility in noise," *J. Acoust. Soc. Am.*, vol. 83, no. 4, pp. 1508–1516, Apr. 1988.

[16] ——, "A clinical test for the assessment of binaural speech perception in noise," *International Journal of Audiology*, vol. 29, no. 5, pp. 275–285, 1990.

[17] ——, "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," *J. Acoust. Soc. Am.*, vol. 92, no. 6, pp. 3132–1319, 1992.

[18] A. W. Bronkhorst, "The cocktail party phenomenon: A review on speech intelligibility in multiple-talker conditions," *Acta Acustica United with Acustica*, vol. 86, no. 1, pp. 117–128, Jan. 2000.

[19] G. Carter, C. Knapp, and A. Nuttall, "Estimation of the magnitude-squared coherence function via overlapped fast Fourier transform processing," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 4, pp. 337–344, Aug. 1973.

[20] A. Chabot-Leclerc, E. N. MacDonald, and T. Dau, "Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain," *J. Acoust. Soc. Am.*, vol. 140, no. 1, pp. 192–205, Jul. 2016.

[21] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomedical Signal Processing and Control*, vol. 8, pp. 311–314, Dec. 2013.

[22] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, Sep. 1953.

[23] J. Clark, C. Yallop, and J. Fletcher, *An Introduction to Phonetics and Phonology*, 3rd ed. Blackwell, 2007, ISBN: 9781405130837.

[24] H. S. Colburn, "Theory of binaural interaction based on auditory-nerve data. II. detection of tones in noise," *J. Acoust. Soc. Am.*, vol. 61, no. 2, pp. 525–533, Feb. 1977.

[25] M. Cooke, "A glimpsing model of speech perception," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, Mar. 2006.

[26] B. Cornelis, M. Moonen, and J. Wouters, "Speech intelligibility improvements with hearing aids using bilateral and binaural adaptive multichannel Wiener filtering based noise reduction," *J. Acoust. Soc. Am.*, vol. 131, no. 6, pp. 4743–4755, 2012.

[27] S. Cosentino, T. Marquardt, D. McAlpine, J. F. Culling, and T. H. Falk, "A model that predicts the binaural advantage to speech intelligibility from the mixed target and interferer signals," *J. Acoust. Soc. Am.*, vol. 135, no. 2, pp. 796–807, Feb. 2014.

[28] J. F. Culling, M. L. Hawley, and R. Y. Litovsky, "The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources," *J. Acoust. Soc. Am.*, vol. 116, no. 2, pp. 1057–1065, Aug. 2004.

[29] ——, "Erratum: The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources [J. Acoust. Soc. Am. 116, 1057 (2004)]," *J. Acoust. Soc. Am.*, vol. 118, no. 1, p. 552, Jul. 2005.

[30] J. A. P. M. de Laat and R. Plomp, "The reception threshold of interrupted speech for hearing-impaired listeners," in *Hearing — Physiological Bases and Psychophysics: Proceedings of the 6th International Symposium on Hearing*, Bad Nauheim, Germany, Apr. 1983.

[31] T. V. den Bogaert, T. J. Klasen, M. Moonen, L. V. Deun, and J. Wouters, "Horizontal localization with bilateral hearing aids: Without is better than with," *J. Acoust. Soc. Am.*, vol. 119, no. 1, pp. 515–526, Jan. 2006.

[32] H. Dillon, *An Introduction to the Physiology of Hearing*, 1st ed. Boomerang Press, 2001, ISBN: 1588900525.

[33] D. D. Dirks and R. H. Wilson, "The effect of spatially separated sound sources on speech intelligibility," *Journal of Speech, Language, and Hearing Research*, vol. 12, pp. 5–38, Mar. 1969.

[34] S. Doclo, T. J. Klasen, T. V. den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions," in *International Workshop Acoust. Echo Noise Control (IWAENC)*, Paris, France, Sep. 2006.

[35] J. R. Dubno, A. R. Horwitz, and J. B. Ahlstrom, "Benefit of modulated maskers for speech recognition by younger and older adults with normal hearing," *J. Acoust. Soc. Am.*, vol. 111, no. 6, pp. 2897–2907, Jun. 2002.

[36] A. J. Duquesnoy, "Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons," *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 739–743, Sep. 1983.

[37] A. J. Duquesnoy and R. Plomp, "Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis," *J. Acoust. Soc. Am.*, vol. 68, no. 2, pp. 537–544, Aug. 1980.

[38] N. I. Durlach, "Equalization and cancellation theory of binaural masking-level differences," *J. Acoust. Soc. Am.*, vol. 35, no. 8, pp. 1206–1218, Aug. 1963.

[39] ——, "Binaural signal detection: Equalization and cancellation theory," in *Foundations of Modern Auditory Theory Volume II*, J. V. Tobias, Ed. New York: Academic Press, 1972, pp. 371–462, ISBN: 9780126919011.

[40] C. Elberling, C. Ludvigsen, and P. E. Lyregaard, "Dantale: A new danish speech material," *Scand. Audiol.*, vol. 18, no. 3, pp. 169–175, 1989.

[41] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Communication*, vol. 41, pp. 331–348, 2003.

[42] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Computational Biology*, vol. 5, no. 3, pp. 1–14, Mar. 2009.

[43] S. Ewert and T. Dau, "Characterizing frequency selectivity for envelope fluctuations," *J. Acoust. Soc. Am.*, vol. 108, no. 3, pp. 1181–1196, Sep. 2000.

[44] T. B. S. Ewert, "Envelope and intensity based prediction of psychoacoustic masking and speech intelligibility," *J. Acoust. Soc. Am.*, vol. 140, no. 2, pp. 1023–1038, Aug. 2016.

[45] T. H. Falk, S. Cosentino, J. Santos, D. Suelzle, and V. Parsa, "Non-intrusive objective speech quality and intelligibility prediction for hearing instruments in complex listening environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vancouver, Canada: IEEE, May 2013.

[46] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[47] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.

[48] J. M. Festen, "Speech-reception threshold in a fluctuating background sound and its possible relation to temporal auditory resolution," in *The Psychophysics of Speech Perception*, M. E. H. Schouten, Ed. Springer, 1987, pp. 461–466, ISBN: 9789400936294.

[49] ——, "Contributions of comodulation masking release and temporal resolution to the speech-reception threshold masked by an interfering voice," *J. Acoust. Soc. Am.*, vol. 94, no. 3, pp. 1295–1300, Sep. 1993.

[50] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acoust. Soc. Am.*, vol. 88, no. 4, pp. 1725–1736, Oct. 1990.

[51] W. T. Fitch, "The evolution of speech: A comparative review," *Trends in Cognitive Sciences*, vol. 4, no. 7, pp. 258–267, Jul. 2000.

[52] ——, *The Evolution of Language*. Cambridge University Press, 2010, ISBN: 9780521677363.

[53] H. Fletcher, "Auditory patterns," *Reviews of Modern Physics*, vol. 12, no. 1, pp. 47–65, Jan. 1940.

[54] H. Fletcher and J. C. Steinberg, "Articulation testing methods," *Bell System Technical Journal*, vol. 8, no. 4, pp. 806–854, Oct. 1929.

[55] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[56] A. C. Furman, S. G. Kujawa, and M. C. Liberman, "Noise-induced cochlear neuropathy is selective for fibers with low spontaneous rates," *Jornal of Neurophysiology*, vol. 110, no. 3, pp. 577–586, Aug. 2013.

[57] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[58] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3679–3689, Dec. 2004.

[59] H. Å. Gustafsson and S. D. Arlinger, "Masking of speech by amplitude-modulated noise," *J. Acoust. Soc. Am.*, vol. 95, no. 1, pp. 518—529, Jan. 1994.

[60] B. Hagermann, "Sentences for testing speech intelligibility in noise," *Scand. Audiol. Suppl.*, vol. 11, no. 2, pp. 79–87, 1982.

[61] V. Hamacher, U. Kornagel, T. Lotter, and H. Puder, "Binaural signal processing in hearing aids: technologies and algorithms," in *Advances in Digital Speech Transmission*, R. Martin, U. Heute, and C. Antweiler, Eds. Wiley, 2008, pp. 401–429.

[62] S. Hochmuth, T. Brand, M. A. Zokoll, F. Z. Castro, N. Wardenga, and B. Kollmeier, "A spanish matrix sentence test for assessing speech reception thresholds in noise," *International Journal of Audiology*, vol. 51, no. 7, pp. 536–544, Jul. 2012.

[63] S. Hochmuth, B. Kollmeier, T. Brand, and T. Jürgens, "Influence of noise type on speech reception thresholds across four languages measured with matrix sentence tests," *International Journal of Audiology*, vol. 54, no. sup2, pp. 62–70, Jun. 2015.

[64] P. Holtegaard and S. Ø. Olsen, "Signs of noise-induced neural degeneration in humans," in *International Symposium on Auditory and Audiological Research (ISAAR)*, Nyborg, Denmark, Aug. 2015, pp. 117–124.

[65] I. Holube, S. Fredelake, M. Vlaming, and B. Kollmeier, "Development and analysis of an international speech test signal (ISTS)," *International Journal of Audiology*, vol. 49, no. 12, pp. 891–903, Dec. 2010.

[66] R. Houben, J. Koopman, H. Luts, K. C. Wagener, A. van Wieringen, H. Verschuure, and W. A. Dreschler, "Development of a Dutch matrix sentence test to assess speech intelligibility in noise," *International Journal of Audiology*, vol. 53, no. 10, pp. 760–763, Oct. 2014.

[67] T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355–367, Jan. 1971.

[68] T. Houtgast, H. J. M. Steeneken, and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function. I. general room acoustics," *Acta Acustica United with Acustica*, vol. 46, no. 1, pp. 60–72, Sep. 1980.

[69] P. A. Howard-Jones and S. Rosen, "Uncomodulated glimpsing in "checkerboard" noise," *J. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2915–2922, May 1993.

[70] ——, "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," *Acta Acustica United with Acustica*, vol. 78, no. 5, pp. 258–272, Jun. 1993.

[71] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall, 2001, ISBN: 9780130226167.

[72] IEEE, "IEEE recommended pratice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, Sep. 1969.

[73] J. C. L. Ingram, *Neurolinguistics: An Introduction to Spoken Language Processing and its Disorders*. Cambridge University Press, 2007, ISBN: 9780521796408.

[74] L. A. Jeffress, "A place theory of sound localization," *J. Acoust. Soc. Am.*, vol. 41, no. 1, pp. 35–39, Mar. 1948.

[75] S. Jelfs, J. F. Culling, and M. Lavandier, "Revision and validation of a binaural model for speech intelligibility in noise," *Hearing Research*, vol. 275, no. 1–2, pp. 96–104, May 2011.

[76] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[77] S. Jørgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acta Acustica United with Acustica*, vol. 101, pp. 1016–1025, 2015.

[78] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Am.*, vol. 130, no. 3, pp. 1475–1487, Sep. 2011.

[79] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 134, no. 1, pp. 436–446, Jul. 2013.

[80] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin-HMM-based non-intrusive speech intelligibility prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China: IEEE, Mar. 2016, pp. 624–628.

[81] J. M. Kates, *Digital Hearing Aids*. Plural Publishing, 2008, ISBN: 9781597563178.

[82] ——, "Improved estimation of frequency importance functions," *J. Acoust. Soc. Am.*, vol. 134, no. 5, pp. EL459–EL464, Nov. 2013.

[83] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, Apr. 2005.

[84] ——, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, Jul. 2014.

[85] D.-S. Kim and A. Tarraf, "ANIQUE+: A new american national standard for non-intrusive estimation of narrowband speech quality," *Bell System Technical Journal*, vol. 12, no. 1, pp. 221–236, 2007.

[86] B. Kollmeier, A. Warzybok, S. Hochmuth, M. A. Zokoll, V. Uslar, T. Brand, and K. C. Wagener, "The multilingual matrix test: Principles, applications, and comparison across languages: A review," *International Journal of Audiology*, vol. 54, no. sup2, pp. 3–16, Sep. 2015.

[87] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Relaxed binaural LCMV beamforming," *IEEE Trans. Audio, Speech, Language Process.*, vol. 25, no. 1, pp. 137–152, Jan. 2017.

[88] K. D. Kryter, "Validation of the Articulation Index," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1698–1702, Nov. 1962.

[89] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1599–1612, May 2016.

[90] A. Kuklasiński and J. Jensen, "Multi-channel Wiener filters in binaural and bilateral hearing aids: speech intelligibility improvement and robustness to DoA error," *Journal of the Audio Engineering Society*, vol. 25, no. 1/2, pp. 8–16, 2017.

[91] P. Ladefoged and I. Maddieson, *The Sounds of the World's Languages*. Brill, 1996, ISBN: 9780631198154.

[92] H. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech," *Journal of Speech, Language, and Hearing Research*, vol. 14, pp. 677–709, Dec. 1971.

[93] M. Lavandier and J. F. Culling, "Prediction of binaural speech intelligibility against noise in rooms," *J. Acoust. Soc. Am.*, vol. 127, no. 1, pp. 387–399, Jan. 2010.

[94] M. Lavandier, S. Jelfs, J. F. Culling, A. J. Watkins, A. P. Raimond, and S. J. Makin, "Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 218–231, Jan. 2012.

[95] T. Leclère, M. Lavandier, and J. F. Culling, "Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking and binaural de-reverberation," *J. Acoust. Soc. Am.*, vol. 137, no. 6, pp. 3335–3345, Jun. 2015.

[96] H. Levitt, "Transformed up-down methods in psychoacoustics," *J. Acoust. Soc. Am.*, vol. 49, no. 2, pp. 467–477, Feb. 1971.

[97] H. Levitt and L. R. Rabiner, "Use of a sequential strategy in intelligibility testing," *J. Acoust. Soc. Am.*, vol. 42, no. 3, pp. 467–477, Sep. 1967.

[98] L. Lightburn and M. Brookes, "SOBM - a binary mask for noisy speech that optimizes an objective intelligibility measure," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Melbourne, Australia: IEEE, Sep. 2015.

[99] ——, "A weighted STOI intelligibility metric based on mutual information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China: IEEE, Mar. 2016, pp. 5365–5369.

[100] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007, ISBN: 9780849350320.

[101] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3387–3405, May 2009.

[102] D. Marquardt, "Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques," Ph.D. dissertation, University of Oldenburg, 2015.

[103] B. T. Meyer, T. Jürgens, T. Wesker, T. Brand, and B. Kollmeier, "Human phoneme recognition depending on speech-intrinsic variability," *J. Acoust. Soc. Am.*, vol. 128, no. 5, pp. 3126—-3141, Nov. 2010.

[104] G. A. Miller, "The masking of speech," *Psychological Bulletin*, vol. 44, no. 2, pp. 105–129, Mar. 1947.

[105] G. A. Miller and J. C. R. Licklider, "The intelligibility of interrupted speech," *J. Acoust. Soc. Am.*, vol. 22, no. 2, pp. 167–173, Mar. 1950.

[106] J. P. Moncur and D. Dirks, "Binaural and monaural speech intelligibility in reverberation," *Journal of Speech, Language, and Hearing Research*, vol. 10, no. 2, pp. 186–195, Jun. 1967.

[107] B. C. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. Brill, 2012, ISBN: 9789004252424.

[108] A. K. Nábělek and J. M. Pickett, "Reception of consonants in a classroom as affected by monaural and binaural listening, noise, reverberation, and hearing aids," *J. Acoust. Soc. Am.*, vol. 56, no. 2, pp. 628–639, Aug. 1974.

[109] P. B. Nelson, S.-H. Jin, A. Carney, and D. Nelson, "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 113, no. 2, pp. 961–968, Feb. 2003.

[110] J. B. Nielsen and T. Dau, "The Danish hearing in noise test," *International Journal of Audiology*, vol. 50, no. 3, pp. 202–208, Mar. 2011.

[111] E. Ozimek, A. Warzybok, and D. Kutzner, "Polish sentence matrix test for speech intelligibility measurement in noise," *International Journal of Audiology*, vol. 49, no. 6, pp. 444–454, Jun. 2010.

[112] E. R. Pedersen and P. M. Juhl, "User-operated speech in noise test: Implementation and comparison with a traditional test," *International Journal of Audiology*, vol. 53, no. 5, pp. 336–344, May 2014.

[113] J. Peissig and B. Kollmeier, "Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners," *Journal of Speech, Language, and Hearing Research*, vol. 101, no. 3, pp. 1660–1670, Jun. 1997.

[114] R. W. Peters, B. C. Moore, and T. Baer, "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 577–587, May 1998.

[115] J. O. Pickles, *An Introduction to the Physiology of Hearing*, 4th ed. Brill, 2013, ISBN: 9789004243774.

[116] C. J. Plack, *The Sense of Hearing*, 2nd ed. Psychology Press, 2014, ISBN: 9781848725157.

[117] R. Plomp and A. M. Mimpen, "Improving the reliability of testing the speech reception threshold for sentences," *Audiology*, vol. 18, no. 1, pp. 43–52, Jan. 1979.

[118] ——, "Effect of the orientation of the speaker's head and the azimuth of a noise source on the speech-reception threshold for sentences," *Acta Acustica United with Acustica*, vol. 48, no. 5, pp. 325–328, Aug. 1981.

[119] G. E. Puglisi, A. Warzybok, S. Hochmuth, C. Visentin, A. Astolfi, N. Prodi, and B. Kollmeier, "An Italian matrix sentence test for the evaluation of speech intelligibility in noise," *International Journal of Audiology*, vol. 54, no. sup2, pp. 44–50, Sep. 2015.

[120] S. R. Quackenbush, *Objective Measures Of Speech Quality*. Prentice Hall, 1988, ISBN: 9780136290568.

[121] A. Radford, M. Atkinson, D. Britain, H. Clahsen, and A. Spencer, *Linguistics: An Introduction*, 2nd ed. Cambridge University Press, 2009, ISBN: 9780521849487.

[122] J. Rennies, T. Brand, and B. Kollmeier, "Prediction of the intelligibility of reverberation on binaural speech intelligibility in noise and in quiet," *J. Acoust. Soc. Am.*, vol. 130, no. 5, pp. 2999–3012, Nov. 2011.

[123] J. Rennies, A. Warzybok, T. Brand, and B. Kollmeier, "Modelling the effects of a single reflection on binaural speech intelligibility," *J. Acoust. Soc. Am.*, vol. 135, no. 3, pp. 1556–1567, Mar. 2014.

[124] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.

[125] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, Dec. 2006.

[126] J. F. Santos and T. H. Falk, "Updating the SRMR-CI metric for improved intelligibility prediction for cochlear implant users," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 2197–2206, Dec. 2014.

[127] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*. Juan les Pins, France: IEEE, Sep. 2014, pp. 55–59.

[128] A. Sarampalis, S. Kalluri, B. Edwards, and E. Hafter, "Objective measures of listening effort: Effects of background noise and noise reduction," *Journal of Speech, Language, and Hearing Research*, vol. 52, pp. 1230–1240, Oct. 2009.

[129] M. R. Schädler, A. Warzybok, S. Hochmuth, and B. Kollmeier, "Matrix sentence intelligibility prediction using an automatic speech recognition system," *International Journal of Audiology*, vol. 54, pp. 1–8, Sep. 2015.

[130] A. Schaub, *Digital Hearing Aids*. Thieme, 2008, ISBN: 9781604060065.

[131] D. Sharma, G. Hilkhuysen, N. D. Gaubitch, P. A. Naylor, M. Brookes, and M. Huckvale, "Data driven method for non-intrusive speech intelligibility estimation," in *European Signal Processing Conference (EUSIPCO)*. Aalborg, Denmark: EURASIP, Aug. 2010, pp. 1899–1903.

[132] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes, "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Communication*, vol. 80, pp. 84–94, 2016.

[133] S. A. Simpson and M. Cooke, "Consonant identification in N-talker babble is a nonmonotonic function of N," *J. Acoust. Soc. Am.*, vol. 118, no. 5, pp. 2775–2778, Nov. 2005.

[134] K. Smeds, A. Leijon, F. Wolters, A. Hammarstedt, S. Båsjö, and S. Hertzman, "Comparison of predictive measures of speech recognition after noise reduction processing," *J. Acoust. Soc. Am.*, vol. 136, no. 3, pp. 1363—-1374, Sep. 2014.

[135] C. Sørensen, J. B. Boldt, F. Gran, and M. G. Christensen, "Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids," in *The European Signal Processing Conference (EUSIPCO)*. Budapest, Hungary: EURASIP, Aug. 2016, pp. 1358–1362.

[136] C. Sørensen, A. Xenaki, J. B. Boldt, and M. G. Christensen, "Pitch-based non-intrusive objective intelligibility prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, United States: IEEE, Mar. 2017, pp. 386–390.

[137] F. Spoor, T. Garland, G. Krovitz, T. M. Ryan, M. T. Silcox, and A. Walker, "The primate semicircular canal system and locomotion," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 26, pp. 10 808—-10 812, May 2007.

[138] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[139] K. N. Stevens, *Acoustic Phonetics*. The MIT Press, 1998, ISBN: 9780262194044.

[140] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[141] ——, "An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech," *J. Acoust. Soc. Am.*, vol. 130, no. 5, pp. 3013–3027, Nov. 2011.

[142] M. M. Taylor and C. D. Creelman, "PEST: Efficient estimates on probability functions," *J. Acoust. Soc. Am.*, vol. 41, no. 4, pp. 782–787, Jan. 1967.

[143] S. van Wijngaarden, "The speech transmission index after four decades of development," *Acoustics Australia*, vol. 40, no. 2, pp. 134–138, Aug. 2012.

[144] S. J. van Wijngaarden and R. Drullman, "Binaural intelligibility prediction based on the speech transmission index," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4514–4523, Jun. 2008.

[145] H. vom Hövel, "Zur bedeutung der übertragungseigenschaften des aussenohrs sowie des binauralen hörsystems bei gestörter sprachübertragung," Ph.D. dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, 1984.

[146] K. Wagener, T. Brand, and B. Kollmeier, "Entwicklung und evaluation eines satztests für die deutsche sprache I: Design des oldenburger satztests," *Zeitschrift für Audiologie*, vol. 38, pp. 4–15, 1999.

[147] ——, "Entwicklung und evaluation eines satztests für die deutsche sprache II: Optimierung des oldenburger satztests," *Zeitschrift für Audiologie*, vol. 38, pp. 44–56, 1999.

[148] ——, "Entwicklung und evaluation eines satztests für die deutsche sprache III: Evaluation des oldenburger satztests," *Zeitschrift für Audiologie*, vol. 38, pp. 86–95, 1999.

[149] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *International Journal of Audiology*, vol. 42, no. 1, pp. 10–17, Jan. 2003.

[150] R. Wan, N. I. Durlach, and H. S. Colburn, "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," *J. Acoust. Soc. Am.*, vol. 128, no. 6, pp. 3678–3690, Sep. 2010.

[151] ——, "Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 768—-776, Aug. 2014.

[152] A. Warzybok, M. Zokoll, N. Wardenga, E. Ozimek, M. Boboshko, and B. Kollmeier, "Development of the Russian matrix sentence test," *International Journal of Audiology*, vol. 54, no. sup2, pp. 35–43, Apr. 2015.

[153] N. Zacharov and S. Bech, *Perceptual Audio Evaluation: Theory, Method and Application*. Wiley-Blackwell, 2006, ISBN: 9780470869239.

[154] M. A. Zokoll, S. Hochmuth, A. Warzybok, K. C. Wagener, M. Buschermöhle, and B. Kollmeier, "Speech-in-noise tests for multilingual hearing screening and diagnostics," *American Journal of Audiology*, vol. 22, no. 1, pp. 175–178, Jun. 2013.

[155] P. M. Zurek, "Binaural advantages and directional effects in speech intelligibility," in *Acoustical factors affecting hearing aid performance*, 2nd ed., G. A. Studebaker and I. Hochberg, Eds. Needham Heights: Allyn and Bacon, 1993, pp. 255–276, ISBN: 0205137784.

Part II

Papers

Paper A

A Binaural Short Time Objective Intelligibility Measure for Noisy and Enhanced Speech

Asger Heidemann Andersen
Jan Mark de Haan
Zheng-Hua Tan
Jesper Jensen

The paper has been presented at
INTERSPEECH, pp. 2563–2567, Dresden, Germany, 2015.

© 2015 ISCA
The layout has been revised.

Abstract

Objective intelligibility measures are increasingly being used to assess the performance of speech processing algorithms, e.g. for hearing aids. It has been shown that the short time objective intelligibility (STOI) measure yields good results in this respect. In this paper we propose a binaural extension of the STOI measure, which predicts binaural advantage using a modified equalization cancellation (EC) stage. The proposed method is evaluated for a range of acoustic conditions. Firstly, the method is able to predict the advantage of spatial separation between a speech target and a speech shaped noise (SSN) interferer. Secondly, the method yields results comparable to the monaural STOI measure when presented with noisy speech processed by ideal time-frequency segregation (ITFS). Finally, the method also performs well when presented with a selection of different acoustic conditions combined with beamforming as used in hearing aids.

1 Introduction

The interest in predicting the intelligibility of noisy speech was initially sparked by the telephone industry where speech intelligibility is of key importance [1]. It is cumbersome and expensive to guide the development in such an industry purely through human listening tests, thus motivating the need for objective means of Speech Intelligibility Prediction (SIP).

The subject of SIP has been widely investigated since the introduction of the Articulation Index (AI) [1], which was later refined and standardized as the Speech Intelligibility Index (SII) [2]. The SII predicts the intelligibility of monaural speech in additive stationary noise. Another early method is the Speech Transmission Index (STI), which focuses on noisy and distorting transmission systems (e.g. telephone lines) [3, 4]. While the SII and STI have been shown to correlate well with human speech intelligibility in some conditions, they work poorly in others. E.g. the methods do not work well when non-linear signal processing is applied to the noisy speech [5, 6]. This has led to the development of a multitude of SIP methods with different characteristics.

One such characteristic is the ability to predict the intelligibility of noise-reduced or enhanced speech [6–12]. The Short-Time Objective Intelligibility (STOI) measure has been successful in predicting the intelligibility of noisy speech which has been processed by Time-Frequency (TF) weighting methods [6], e.g. as used in noise reduction systems.

Another characteristic is the ability to predict the advantage obtained through binaural listening. It is well known that humans may obtain a great advantage in conditions where the speech and noise sources are spatially separated [13]. This effect can be predicted by the Equalization Cancellation (EC)

model [14, 15], which has been used in the development of binaural SIP methods. The Binaural Speech Intelligibility Measure (BSIM) is essentially the combination of the EC model with the SII [16, 17]. Another EC-based binaural SIP method is proposed in [18–20], which focuses on predicting binaural advantage for different room acoustical conditions. Both of these methods, however, perform SIP for the case of additive noise and no processing.

These characteristics are both highly relevant in many areas, e.g. for the development of hearing aids and cochlear implants [12, 21]. Modern hearing aids apply non-linear speech enhancement algorithms and may modify the binaural cues presented to the user. Hence, binaural SIP of enhanced speech could be used to guide developments in this field. Unfortunately, to the knowledge of the authors, no existing SIP method can handle binaural processed signals (e.g. predict the effect that a speech enhancement algorithm has on binaural advantage).

In this paper, we propose a binaural SIP method which is able to predict the intelligibility of noisy, potentially processed, speech. Specifically, we combine a modified EC stage with the STOI measure to obtain such a method. We refer to the proposed method as the Binaural STOI (BSTOI) measure. The method is presented and evaluated for a range of conditions including Ideal Time Frequency Segregation (ITFS) and spatial separation between target and interferer.

2 The BSTOI Method

The objective of the BSTOI measure is to predict the intelligibility of binaural speech which may be corrupted by additive noise and processed by a speech enhancement algorithm. The method is intrusive in the sense that it assumes knowledge of the clean speech signal as a reference. The method therefore takes four inputs: the clean speech as recorded at the left and right ears of the listener, and the corresponding noisy/processed speech at the left and right ears. As output, the method produces an index in the range of 0 to 1, which should correlate well with the fraction of the spoken words which are understandable to a normal hearing listener. The BSTOI method is outlined by the block diagram in Fig. A.1.

2.1 Step 1: TF Decomposition

The first step of the method is to perform a short time DFT-based TF decomposition of the four input signals. This is done in exactly the same manner as in the STOI-method [6]. Let $\hat{x}_{k,m}^{(l)} \in \mathbb{C}$ be the TF unit corresponding to the clean signal at the left ear at the m'th time frame and the k'th frequency bin. Similarly, let $\hat{x}_{k,m}^{(r)}$, $\hat{y}_{k,m}^{(l)}$ and $\hat{y}_{k,m}^{(r)}$ denote the right ear clean signal and the

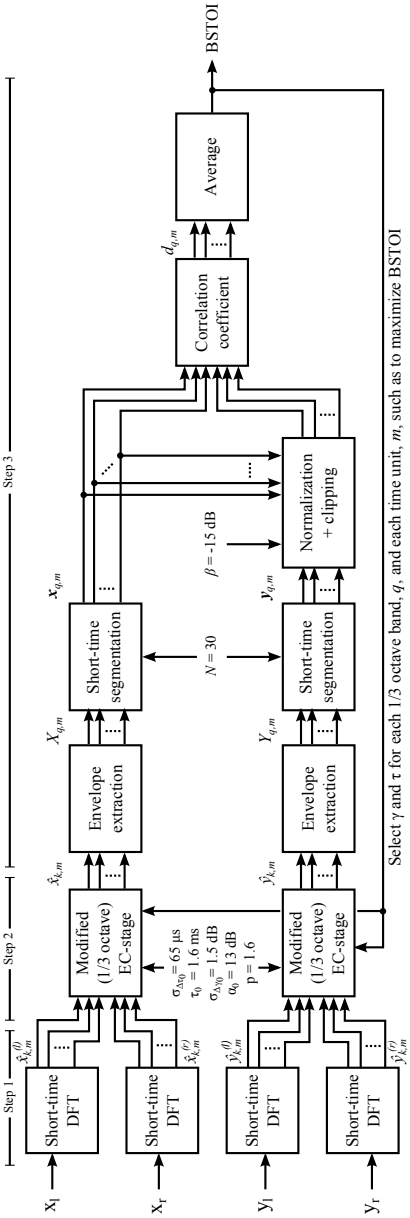


Fig. A.1: A block diagram which illustrates the computation of the BSTOI measure. The measure has the option of computing the intermediate correlation coefficients, $d_{q,m}$, in a better-ear fashion whenever favourable. This is omitted from the figure for simplicity.

processed left and right ear signal TF units.

2.2 Step 2: EC Processing

The second step of the method combines the left and right ear signals using a modified EC stage to model binaural advantage. The original EC stage assumes four inputs: left/right ear clean signal (e.g. speech) and left/right ear noise [14, 15]. The method time shifts and amplitude adjusts the left and right ear signals relative to one-another (equalization) and then proceeds to subtract the left ear signals from the right ear signals (cancellation) to obtain two output signals: one clean difference signal and one noise difference signal. The level of time shift and amplitude adjustment between the left and right ears is determined such that the Signal to Noise Ratio (SNR) at the output is maximized. The BSTOI method assumes access to the clean input signals but the noise is not assumed separately available. Instead, a potentially processed combination of speech and noise is available. The SNR is not directly computable from these signals, and therefore a conventional EC stage cannot be applied. Instead, we modify the EC stage by introducing two particular changes:

1. The left/right clean signals and the left/right *processed/noisy signals* are combined in exactly the same manner as the left/right clean signals and the left/right *noise signals* are combined in the conventional EC stage.

2. Instead of relying on maximizing the SNR for determining the relative time and level adjustments between the ears, the STOI measure of the resulting clean and processed signals is maximized.

The second modification is inspired by [16, 17] who found time and level adjustments in the EC stage, that maximized the SII. The linear combination of left and right ear clean signals in the EC stage is described by:

$$\hat{x}_{k,m} = \lambda\, \hat{x}^{(l)}_{k,m} - \lambda^{-1} \hat{x}^{(r)}_{k,m}, \tag{A.1}$$

where:

$$\lambda = 10^{(\gamma+\Delta\gamma)/40} e^{j\omega(\tau+\Delta\tau)/2}. \tag{A.2}$$

The factor λ implements the equalization step in the EC stage. The left and right ear signals are level adjusted by γ (in dB) and time shifted by τ relative to one-another and are thereafter subtracted. The processed signals are treated similarly, to obtain a combined TF unit $\hat{y}_{k,m}$. To obtain performance similar to that of humans, the EC stage adds noise sources, $\Delta\gamma$ and $\Delta\tau$, to the values of γ and τ [14, 15, 22]. These noise sources are normally distributed

with zero mean and standard deviation (adapted from [22] in the same manner as is done in [16, 17])[1]:

$$\sigma_{\Delta\gamma} = \sqrt{2} \cdot \sigma_{\Delta\gamma_0} \left(1 + \left(\frac{|\gamma|}{\alpha_0}\right)^p\right), \tag{A.3}$$

$$\sigma_{\Delta\tau} = \sqrt{2} \cdot \sigma_{\Delta\tau_0} \left(1 + \frac{|\tau|}{\tau_0}\right), \tag{A.4}$$

with $\sigma_{\Delta\gamma_0} = 1.5$ dB, $\alpha_0 = 13$ dB, $p = 1.6$, $\sigma_{\Delta\tau_0} = 65$ μs and $\tau_0 = 1.6$ ms. The determination of the values γ and τ is covered in Section 2.4.

2.3 Step 3: Intelligibility Prediction

At this point, the method progresses like the STOI method. The clean and processed signal envelopes are determined in $Q = 15$ third octave bands [6]:

$$X_{q,m} = \sqrt{\sum_{k=k_1(q)}^{k_2(q)-1} \left|\hat{x}_{k,m}\right|^2}, \tag{A.5}$$

where $k_1(q)$ and $k_2(q)$ denote the lower and upper DFT bins for the q'th third octave band. The same is done for the processed signal to obtain third octave envelopes, $Y_{q,m}$. These envelopes are arranged into vectors of $N = 30$ samples [6]:

$$\mathbf{x}_{q,m} = [X_{q,m-N+1},\ X_{q,m-N+2},\ \ldots,\ X_{q,m}]^\mathsf{T}. \tag{A.6}$$

Similar vectors, $\mathbf{y}_{q,m} \in \mathbb{R}^{N\times 1}$ are defined for the processed signal. The processed envelope is subjected to a clipping procedure which bounds the sensitivity of the model to severely degraded frames (see [6] for details):

$$\bar{\mathbf{y}}(n) = \min\left(\frac{||\mathbf{x}_{q,m}||}{||\mathbf{y}_{q,m}||}\mathbf{y}_{q,m}(n),\ \ (1 + 10^{-\beta/20})\mathbf{x}_{q,m}(n)\right), \tag{A.7}$$

with $\beta = -15$ (dB) and $n = 1,\ 2,\ \ldots,\ N$. Each clean signal envelope is then correlated with the corresponding clipped processed signal envelope [6]:

$$d_{q,m} = \frac{(\mathbf{x}_{q,m} - \mu_{\mathbf{x}_{q,m}})^\mathsf{T}(\bar{\mathbf{y}}_{q,m} - \mu_{\bar{\mathbf{y}}_{q,m}})}{||\mathbf{x}_{q,m} - \mu_{\mathbf{x}_{q,m}}||||\bar{\mathbf{y}}_{q,m} - \mu_{\bar{\mathbf{y}}_{q,m}}||}, \tag{A.8}$$

where $\mu_{(\cdot)}$ denotes the mean of the entries in the corresponding vector. The final BSTOI measure is obtained by averaging these intermediate correlation

[1] In [22], noise is added separately to the left and right ear signals. Here, one noise source is applied symmetrically. This leads to the addition of a factor of $\sqrt{2}$ in (A.3) and (A.4) compared to [22].

coefficients [6]:

$$\text{BSTOI} = \frac{1}{QM} \sum_{q=1}^{Q} \sum_{m=1}^{M} d_{q,m}, \tag{A.9}$$

where Q and M is the number of frequency bands and the number of frames, respectively.

2.4 Determination of γ and τ

Finally, we consider the determination of the parameters γ and τ. As stated, these are determined such as to maximize the final BSTOI measure. The parameters are determined individually for each time unit, m, and third octave band, q. Thus, each intermediate correlation coefficient is a function of its own set of parameters, $d_{q,m} = d_{q,m}(\gamma, \tau)$. The BSTOI measure, (A.9), can therefore be maximized by maximizing each of the intermediate correlation coefficients individually:

$$d_{q,m} = \max_{\gamma,\tau} \quad d_{q,m}(\gamma, \tau). \tag{A.10}$$

In practice, $d_{q,m}$ is evaluated for a discrete set of γ and τ values (a "grid") and the highest value is chosen.

It should be noted that the BSTOI measure is stochastic due to the noise sources, $\Delta\gamma$ and $\Delta\tau$, in (A.2). The approximately optimal values of γ and τ are therefore first determined without the noise. Using these γ and τ values, the measure is averaged over ten noise realizations. In addition, the intermediate correlation coefficients are computed for the left and right ears separately, with the monaural STOI method [6]. Whenever a single ear provides a higher correlation than the corresponding EC processed alternative, the better-ear correlation is used.

3 Evaluation

Since the BSTOI measure accepts the combination of binaural and processed input signals, it is applicable to a wide range of acoustical conditions. Here, we evaluate the method against results from three different listening tests spanning different types of conditions.

3.1 S_0N_θ with Dantale Target and SSN Interferer

We first investigate predictions of the binaural advantage in the case of a frontal speech target and a single Speech Shaped Noise (SSN) interferer from

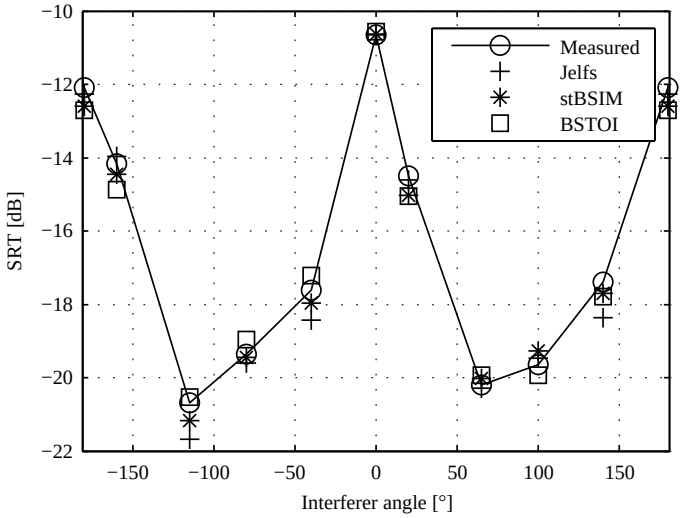


Fig. A.2: Measured and predicted SRTs for a frontal target and ten different interferer angles. Measured data points are connected by lines for ease of viewing.

different directions, θ, in the horizontal plane under anechoic conditions (denoted S_0N_θ) [13, 18]. Sentences from the Danish Dantale II corpus [23] were used as target material, while SSN was generated by filtering Gaussian noise to obtain the same long-time average spectrum as that of the speech material. Both speech and noise was filtered with large pinnae KEMAR Head Related Transfer Functions (HRTFs) from the CIPIC database [24] to simulate the required binaural conditions. The signals were presented to the subjects by headphones. Ten normal hearing subjects participated in the listening experiment. Each subject was presented with sentences in noise from ten different angles and six fixed SNRs. The subjects were instructed to repeat any words they heard, and the number of correct words were recorded for each presented sentence. This resulted in the scoring of (10 subjects)$\times$(10 interferer angles)$\times$(6 SNRs)$\times$(3 repetitions) = 1800 sentences. The resulting data was averaged across all subjects and repetitions to obtain a total of 60 average scores. Logistic functions were fitted to data from each interferer angle, and 50% Speech Reception Thresholds (SRTs)[2] were estimated from these.

Fig. A.2 compares the measured SRTs with the predictions of three different binaural SIP methods, including the BSTOI measure. The method denoted as Jelfs was introduced in [19] and an implementation from the Auditory Modelling Toolbox [25] was used. The method predicts only relative SRTs and the predictions have therefore been offset such that the pre-

[2]The 50% SRT is the SNR where the subject scores 50% correct words.

dictions match the measured results exactly in the S_0N_0-condition. The stBSIM method was implemented as best as possible according to the description given in [17]. The Jelfs and stBSIM methods were both supplied with frontal SSN as target and SSN from different directions as interferer (to most closely resemble the conditions under which these methods have previously been validated [17, 19]).

The BSTOI was supplied with an input signal consisting of 30 (randomly chosen) concatenated Dantale II sentences as speech and binaural SSN as interferer. The processed signal input to the BSTOI measure was generated simply as the sum of clean speech and interferer (as no processing was carried out). SRTs were determined with the BSTOI measure in a manner similar to that used in [17, 26, 27]: The BSTOI output at the measured SRT in the S_0N_0-condition was computed and used as a reference. The SRTs of other conditions were then predicted by adaptively varying the input SNR until this reference value was found as the output of the BSTOI measure.

Fig. A.2 shows that BSTOI is able to predict the binaural advantage rather accurately. The largest prediction error is just under 0.7 dB. The performance of BSTOI is similar to that of Jelfs and stBSIM. The results suggest that the changes introduced to the EC model, to merge it with the monaural STOI measure, do not significantly impair its performance under the studied conditions. Furthermore, they indicate that the use of STOI in conjunction with the EC model is a viable approach to predicting binaural speech intelligibility.

3.2 Ideal Time Frequency Segregation

The STOI measure has shown to perform well at predicting the intelligibility of speech which has been processed by ITFS [6]. It is obviously desired that this property is retained in the binaural extension of the measure. We therefore applied the BSTOI measure to the same collection of monaural ITFS data as used in [6]. The data presented in [28] consist of clean and noisy/processed speech signals from the Dantale II corpus as well as the results from a listening test with 15 normal hearing subjects. The signals are contaminated by four noise types: SSN as well as cafeteria, bottling factory and car interior noise [28]. The processing consist of ITFS with two different mask types: Ideal Binary Masks (IBMs) and Target Binary Masks (TBMs) [28]. The listening test has been carried out in a manner quite similar to that described in the previous section (see [28] for details). The listening test results were averaged across all subjects and repetitions for each condition. Each condition was scored by the STOI and BSTOI measures using 30 concatenated Dantale II sentences as input. The signals were presented diotically to the BSTOI measure (the same signals were given for the left and right ear inputs).

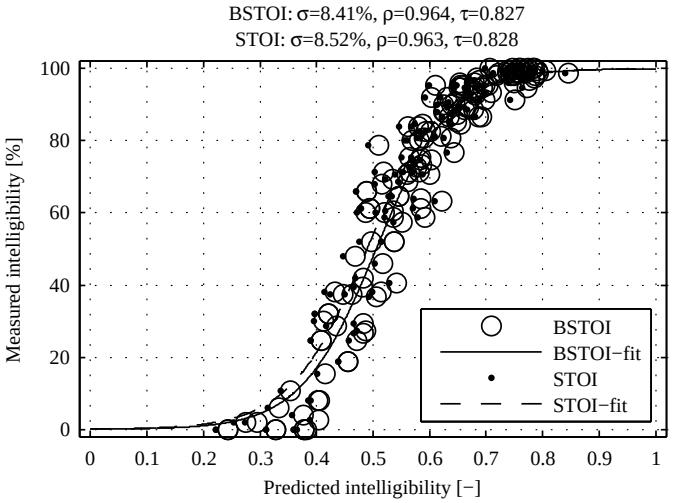


Fig. A.3: Measured intelligibility (averaged over 15 test subjects and multiple sentence scores) vs. STOI and BSTOI measures, and fitted logistic curves. The values above the plot show sample standard deviation from the respective logistic curves (σ), correlation coefficient (ρ), and Kendall's tau (τ).

Fig. A.3 shows the resulting scores vs. the average listening test scores. The results for the STOI measure are close to those presented in [6] as should be expected. Slight variations occur from the use of different Dantale II sentences as input. The BSTOI measure shows predictions which are similar to those of STOI albeit slightly higher. This relative difference is not important as none of the measures predict intelligibility relative to any fixed reference. A number of key measures are shown on the figure as well. These all indicate that the BSTOI measure performs similar to the STOI measure for this dataset. This is as expected, as there is no advantage to be gained by the addition of the EC stage in the diotic condition. On the other hand, the results indicate that extending STOI with a modified EC stage does not negatively affect its ability to predict the intelligibility of ITFS processed speech.

3.3 Various Conditions with and without Beamforming

Lastly, we evaluate the BSTOI measure for a selection of different conditions. A listening experiment was carried out in a manner almost identical to that discussed in Section 3.1. Ten normal hearing subjects were presented with Dantale II sentences in six different conditions for a scoring of (10 subjects)$\times$(6 conditions)$\times$(6 SNRs)$\times$(3 repetitions) = 1080 sentences in total. These were averaged across subjects and repetitions to produce a total of 36 data points. All six conditions were anechoic with speech originating from the front. The first condition was contaminated by isotropic ("Iso") SSN.

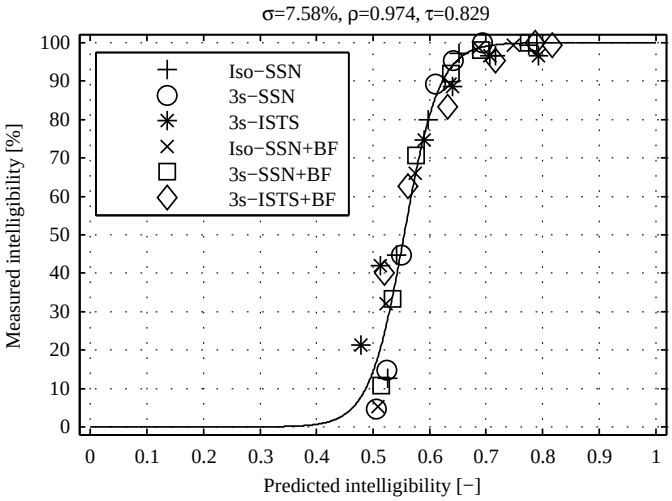


Fig. A.4: The averged results of the third discussed listening test vs. the averaged BSTOI predictions. A fitted logistic function is shown as well.

The second condition was contaminated by uncorrelated SSN from point sources at 110°, 180° and −110° in the horizontal plane ("3s"). The third condition considered the same layout of noise sources as condition two, but used three different segments of the International Speech Test Signal (ISTS) as noise. The ISTS is a speech-like noise signal created from recorded speech, but which is largely non-intelligible [29]. Conditions four to six are the same as one to three, but include monaural 2-microphone Minimum Variance Distortionless Response (MVDR) beamforming as used in a behind-the-ear hearing aid ("BF"). The signals which were presented to the subjects were saved. BSTOI predictions were made for each of these signals and averaged in the same way as the measured data.

Fig. A.4 shows the results. While the investigated conditions are highly diverse, the BSTOI predictions appear to be almost monotonically related to the measured intelligibility. At the point of low measured intelligibility, it appears that the method slightly underestimates the intelligibility in the ISTS conditions relative to the SSN conditions.

4 Conclusions

In this paper we present a binaural extension, BSTOI, of the Short-Time Objective Intelligibility (STOI) measure. The method is based on extending the STOI measure using a modified Equalization Cancellation (EC) stage to model binaural advantage. Unlike existing methods, BSTOI is able to pre-

dict intelligibility for signals which are binaural *and* have been processed by speech enhancement algorithms. Initial experiments show promising results. Firstly, they indicate that BSTOI can predict both 1) the binaural advantage of spatial separation between a target and an interferer and 2) the intelligibility of ITFS processed speech presented diotically. Secondly, BSTOI predicts intelligibility well in a variety of different conditions with and without linear processing (beamforming). The investigated conditions, however, cover only a small part of the domain of input signals, processing methods and acoustical conditions to which the measure is theoretically applicable, and the evaluation can therefore only be considered as an initial validation. Future work includes further investigation of the proposed method for more complicated combinations of spatial layouts and binaural non-linear processing.

5 Acknowledgements

This work was funded by the Oticon Foundation and the Danish Innovation Foundation.

References

[1] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[2] *Methods for Calculation of the Speech Intelligibility Index*, American National Standards Institute Std. S3.5-1997, 1997.

[3] T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355–367, Jan. 1971.

[4] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[5] C. Ludvigsen, C. Elberling, , and G. Keidser, "Evaluation of a noise reduction method – comparison between observed scores and scores predicted from STI," *Scand. Audiol. Suppl.*, vol. 38, pp. 50–55, 1993.

[6] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for inteligibility prediction of time-frequency weighted noisy speech," *IEEE Tran. on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[7] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3679–3689, Dec. 2004.

[8] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, Apr. 2005.

[9] J. B. Boldt and D. P. W. Ellis, "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in *Proceedings of the 17th European Signal Processing Conference (EUSIPCO-2009)*. Glasgow, Scotland: EURASIP, Aug. 2009.

[10] C. Christiansen, M. S. Pedersen, and T. Dau, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Communication*, vol. 52, no. 7-8, pp. 678–692, Jul. 2010.

[11] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE Tran. on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 430–440, Feb. 2014.

[12] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[13] A. W. Bronkhorst, "The cocktail party phenomenon: A review on speech intelligibility in multiple-talker conditions," *Acta Acustica United with Acustica*, vol. 86, no. 1, pp. 117–128, Jan. 2000.

[14] N. I. Durlach, "Equalization and cancellation theory of binaural masking-level differences," *J. Acoust. Soc. Am.*, vol. 35, no. 8, pp. 1206–1218, Aug. 1963.

[15] ——, "Binaural signal detection: Equalization and cancellation theory," in *Foundations of Modern Auditory Theory Volume II*, J. V. Tobias, Ed. New York: Academic Press, 1972, pp. 371–462.

[16] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 331–342, Apr. 2006.

[17] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension and evaluation of a binaural speech intelligibility model," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2479–2497, Dec. 2010.

[18] M. Lavandier and J. F. Culling, "Prediction of binaural speech intelligibility against noise in rooms," *J. Acoust. Soc. Am.*, vol. 127, no. 1, pp. 387–399, Jan. 2010.

[19] S. Jelfs, J. F. Culling, and M. Lavandier, "Revision and validation of a binaural model for speech intelligibility in noise," *Hearing Research*, vol. 275, no. 1–2, pp. 96–104, May 2011.

[20] M. Lavandier, S. Jelfs, J. F. Culling, A. J. Watkins, A. P. Raimond, and S. J. Makin, "Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 218–231, Jan. 2012.

[21] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Matching pursuit for channel selection in coclear implants based on an intelligibility metric," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO-2012)*. Bucharest, Romania: EURASIP, Aug. 2012.

[22] H. vom Hövel, "Zur bedeutung der übertragungseigenschaften des aussenohrs sowie des binauralen hörsystems bei gestörter sprachübertragung," Ph.D. dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, 1984.

[23] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *International Journal of Audiology*, vol. 42, no. 1, pp. 10–17, Jan. 2003.

[24] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proccedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct. 2001.

[25] P. Søndergaard and P. Majdak, "The auditory modeling toolbox," in *The Technology of Binaural Listening*, ser. Modern Acoustics and Signal Processing, J. Blauert, Ed. Springer, 2013, pp. 33–56.

[26] R. Wan, N. I. Durlach, and H. S. Colburn, "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," *J. Acoust. Soc. Am.*, vol. 128, no. 6, pp. 3678–3690, Sep. 2010.

[27] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.

[28] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.

[29] I. Holube, S. Fredelake, M. Vlaming, and B. Kollmeier, "Development and analysis of an international speech test signal (ISTS)," *International Journal of Audiology*, vol. 49, no. 12, pp. 891–903, Dec. 2010.

Paper B

A Method for Predicting the Intelligibility of Noisy and Non-Linearly Enhanced Binaural Speech

Asger Heidemann Andersen
Jan Mark de Haan
Zheng-Hua Tan
Jesper Jensen

The paper has been presented at the
41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, pp. 4995-4999, 2016.

© 2016 IEEE
The layout has been revised.

Abstract

We propose and evaluate a binaural speech intelligibility measure. The measure is a binaural extension of the Short-Time Objective Intelligibility (STOI) measure and focuses on predicting the intelligibility of noisy speech which has been enhanced by a speech processing algorithm (e.g. in a hearing aid). We show that the measure can accurately predict 1) the Speech Reception Threshold (SRT) for a frontal speaker masked by a point noise source in the horizontal plane, 2) the improvement in SRT obtained by independently processing the left and right ear signals with Ideal Time Frequency Segregation (ITFS), and 3) the intelligibility of speech in the presence of multiple interferers as well as the effect of processing the noisy signals with 2-microphone MVDR beamforming as used in hearing aids. Finally, we show that the computational demands associated with the measure are favourable in comparison with those of a previously proposed measure with similar properties.

1 Introduction

The topic of Speech Intelligibility Prediction (SIP) has been widely investigated since the introduction of the Articulation Index (AI) [1], which was later refined and standardized as the Speech Intelligibility Index (SII) [2]. While the research interest initially came from the telephone industry [1], the possible application to hearing aids and cochlear implants has recently gained attention [3, 4].

The SII predicts monaural intelligibility in conditions with additive, stationary noise. Another early and highly popular method is the Speech Transmission Index (STI), which predicts the intelligibility of speech, which has been transmitted through a noisy and distorting transmission system (e.g. a reverberant room) [5, 6]. Many additional SIP methods have been proposed, mainly with the purpose of extending the range of conditions under which predictions can be made (e.g. [7–17]).

For SIP methods to be applicable in relation to binaural communication devices such as hearing aids, the operating range of the classical methods must be expanded in two ways. Firstly, they must be able to take into account the non-linear processing that typically happens in such devices. This task is complicated by the fact that many SIP methods assume knowledge of the clean speech and interferer in separation; an assumption which is not meaningful when the combination of speech and noise has been processed non-linearly. One example of a method which does not make this assumption, is the STOI measure [7] which predicts intelligibility from a noisy/processed signal and a clean speech signal. The STOI measure has been shown to predict well the influence on intelligibility of multiple enhancement algorithms [7]. Secondly, SIP methods must take into account the

fact that signals are commonly presented binaurally to the user. Binaural auditory perception provides the user with different degrees of advantage, depending on the acoustical conditions and the applied processing [18]. Several SIP methods have focused on predicting this advantage [11–17]. Existing binaural methods, however, can generally not provide predictions for non-linearly processed signals.

In [19] we proposed a binaural extension of the STOI measure: the Binaural STOI (BSTOI) measure. The BSTOI measure was shown to predict well the intelligibility (including binaural advantage) obtained in conditions with a frontal target and a single point noise source in the horizontal plane. The BSTOI measure was also shown to predict the intelligibility of diotic speech which had been processed by ITFS.

In this paper we present an improved version of the BSTOI measure which is computationally less demanding and, unlike BSTOI, produces deterministic results. We furthermore show that the proposed measure is able to predict intelligibility in conditions where *both* binaural advantage *and* non-linear processing simultaneously influence intelligibility. To the knowledge of the authors, no other SIP method is capable of producing predictions in conditions where intelligibility is affected by both. We refer to the improved binaural speech intelligibility measure as the Deterministic BSTOI (DBSTOI) measure.

2 The DBSTOI Measure

The DBSTOI measure scores intelligibility based on four signals: The noisy/-processed signal as presented to the left and right ears of the listener and a clean speech signal, also at both ears. The clean signal should be the same as the noisy/processed one, but with neither noise or processing. The DBSTOI measure produces a score in the range 0 to 1. The aim is to have a monotonic correspondence between the DBSTOI measure and measured intelligibility, such that a higher DBSTOI measure corresponds to a higher intelligibility (e.g. percentage of words heard correctly).

The DBSTOI measure is based on combining a modified Equalization Cancellation (EC) stage with the STOI measure as proposed in [19]. Here, we introduce further structural changes in the STOI measure to allow for better integration with the EC-stage. This allows for computing the measure deterministically and in closed form, contrary to the BSTOI measure [19], which is computed using Monte Carlo simulation.

The structure of the DBSTOI measure is shown in Fig. B.1. The procedure is separated in three main steps: 1) a time-frequency-decomposition based on the Discrete Fourier Transformation (DFT), 2) a modified EC stage which extracts binaural advantage and 3) a modified version of the monaural STOI measure. The three steps are described in Secs. 2.1, 2.2 and 2.3, respectively.

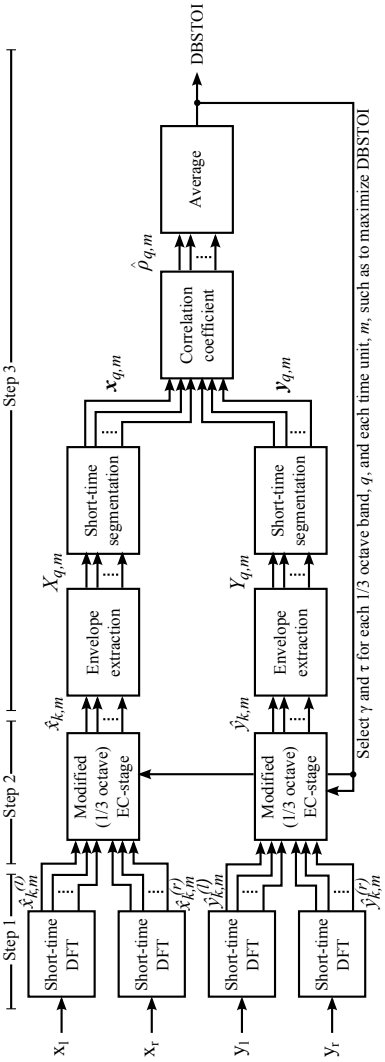


Fig. B.1: A block diagram which illustrates the computation of the DBSTOI measure.

2.1 Step 1: TF Decomposition

The first step resamples the four input signals to 10 kHz, removes segments with no speech (via an ideal frame based voice activity detector) and performs a short-time DFT-based Time-Frequency (TF) decomposition. This is done in exactly the same manner as for the STOI measure [7]. Let $\hat{x}_{k,m}^{(l)} \in \mathbb{C}$ be the TF unit corresponding to the clean signal at the left ear at the m'th time frame and the k'th frequency bin. Similarly, let $\hat{x}_{k,m}^{(r)}$, $\hat{y}_{k,m}^{(l)}$ and $\hat{y}_{k,m}^{(r)}$ denote the right ear clean signal and the left and right ear processed signal TF units, respectively.

2.2 Step 2: EC Processing

The second step of computing the measure combines the left and right ear signals using a modified EC stage to model binaural advantage [20, 21].

A combined clean signal is obtained by relatively time shifting and amplitude adjusting the left and right clean signals and thereafter subtracting one from the other. The same is done for the noisy/processed signals to obtain a single noisy/processed signal. The relative time shift of τ (seconds) and amplitude adjustment of γ (dB) is given by the factor:

$$\lambda = 10^{(\gamma+\Delta\gamma)/40} e^{j\omega(\tau+\Delta\tau)/2}, \tag{B.1}$$

where $\Delta\tau$ and $\Delta\gamma$ are uncorrelated noise sources which model imperfections of the human auditory system [20–22]. The resulting combined clean signal is given by:

$$\hat{x}_{k,m} = \lambda\, \hat{x}_{k,m}^{(l)} - \lambda^{-1} \hat{x}_{k,m}^{(r)}. \tag{B.2}$$

A combined noisy/processed TF-unit, $\hat{y}_{k,m}$, is obtained in a similar manner (using the same value of λ).

The uncorrelated noise sources, $\Delta\tau$ and $\Delta\gamma$, are normally distributed with zero mean and standard deviation (adapted from [22] in the same manner as is done in [11, 12])[1]:

$$\sigma_{\Delta\gamma}(\gamma) = \sqrt{2} \cdot 1.5 \text{ dB} \cdot \left(1 + \left(\frac{|\gamma|}{13 \text{ dB}}\right)^{1.6}\right), \qquad [\text{dB}] \tag{B.3}$$

$$\sigma_{\Delta\tau}(\tau) = \sqrt{2} \cdot 65 \cdot 10^{-6} \text{ s} \cdot \left(1 + \frac{|\tau|}{0.0016 \text{ s}}\right). \qquad [\text{s}] \tag{B.4}$$

Following the principle introduced in [19], the values γ and τ are determined such as to maximize the scoring of intelligibility. This is covered in Sec. 2.4.

[1] In [22], noise is added separately to the left and right ear signals. Here, one noise source is applied symmetrically. This leads to a multiplicative factor of $\sqrt{2}$ in (B.3) and (B.4) compared to [22].

2.3 Step 3: Intelligibility Prediction

At this point the four input signals have been condensed to two signals: a clean signal, $\hat{x}_{k,m}$, and a noisy/processed signal, $\hat{y}_{k,m}$. We compute an intelligibility score for these signals by use of a variation of the STOI measure[2].

The clean and processed signal power envelope is determined in $Q = 15$ third octave bands:

$$\begin{aligned} X_{q,m} &= \sum_{k=k_1(q)}^{k_2(q)} \left|\hat{x}_{k,m}\right|^2 \\ &\approx \alpha X_{q,m}^{(l)} + \alpha^{-1} X_{q,m}^{(r)} - 2\,\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)} X_{q,m}^{(c)}\right], \end{aligned} \tag{B.5}$$

where $\alpha = 10^{\frac{\gamma+\Delta\gamma}{20}}$ and:

$$X_{q,m}^{(l)/(r)} = \sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m}^{(l)/(r)}|^2, \qquad X_{q,m}^{(c)} = \sum_{k=k_1(q)}^{k_2(q)} \hat{x}_{k,m}^{(l)*} \hat{x}_{k,m}^{(r)}, \tag{B.6}$$

and where $k_1(q)$ and $k_2(q)$ denote the lower and upper DFT bins for the q'th third octave band, respectively, and ω_q is the center frequency of the q'th frequency band. The approximate equality is obtained by inserting (B.1) and (B.2) and assuming that the energy in each third octave band is contained at the center frequency. A similar procedure for the processed signal yields third octave power envelopes, $Y_{q,m}$.

If we assume that the input signals are wide sense stationary stochastic processes, the power envelopes, $X_{q,m}$ and $Y_{q,m}$ are also stochastic processes, due to the stochastic nature of the input signals as well as the noise sources, $\Delta\tau$ and $\Delta\gamma$, in the EC stage. An underlying assumption of STOI is that intelligibility is related to the correlation between clean and noisy/-processed envelopes [7]:

$$\rho_q = \frac{E\left[\left(X_{q,m} - E\left[X_{q,m}\right]\right)\left(Y_{q,m} - E\left[Y_{q,m}\right]\right)\right]}{\sqrt{E\left[\left(X_{q,m} - E\left[X_{q,m}\right]\right)^2\right] E\left[\left(Y_{q,m} - E\left[Y_{q,m}\right]\right)^2\right]}}, \tag{B.7}$$

where the expectation is taken across both input signals and the noise sources in the EC stage. To estimate ρ_q, the power envelopes are arranged into vectors of $N = 30$ samples [7]:

$$\mathbf{x}_{q,m} = [X_{q,m-N+1},\ X_{q,m-N+2},\ \ldots,\ X_{q,m}]^\mathsf{T}. \tag{B.8}$$

[2]For mathematical tractability, we use power envelopes rather than magnitude envelopes as originally proposed in STOI [7]. This is also done in [3] and appears not to have a significant effect on predictions [3, 23]. Furthermore, we discard the clipping mechanism contained in the original STOI, as also done in [3]. We have seen no indication that this negatively influences results.

Similar vectors, $\mathbf{y}_{q,m} \in \mathbb{R}^{N\times 1}$ are defined for the processed signal. An N-sample estimate of ρ_q across the input signals is then given by:

$$\hat{\rho}_{q,m} = \frac{E_\Delta\left[(\mathbf{x}_{q,m} - \mathbf{1}\mu_{\mathbf{x}_{q,m}})^\intercal(\mathbf{y}_{q,m} - \mathbf{1}\mu_{\mathbf{y}_{q,m}})\right]}{\sqrt{E_\Delta\left[||\mathbf{x}_{q,m} - \mathbf{1}\mu_{\mathbf{x}_{q,m}}||^2\right] E_\Delta\left[||\mathbf{y}_{q,m} - \mathbf{1}\mu_{\mathbf{y}_{q,m}}||^2\right]}}, \tag{B.9}$$

where $\mu_{(\cdot)}$ denotes the mean of the entries in the given vector, E_Δ is the expectation across the noise in the EC stage and $\mathbf{1}$ is the vector of all ones. A closed form expression for this expectation can be derived, and is given by (derivation omitted):

$$\begin{aligned}
&E_\Delta\left[(\mathbf{x}_{q,m} - \mathbf{1}\mu_{\mathbf{x}_{q,m}})^\intercal(\mathbf{y}_{q,m} - \mathbf{1}\mu_{\mathbf{y}_{q,m}})\right] = \\
&\quad(e^{2\beta}\mathbf{l}_{x_{q,m}}^\intercal\mathbf{l}_{y_{q,m}} + e^{-2\beta}\mathbf{r}_{x_{q,m}}^\intercal\mathbf{r}_{y_{q,m}})e^{2\sigma_{\Delta\beta}^2} \\
&+ \mathbf{r}_{x_{q,m}}^\intercal\mathbf{l}_{y_{q,m}} + \mathbf{l}_{x_{q,m}}^\intercal\mathbf{r}_{y_{q,m}} - 2e^{\sigma_{\Delta\beta}^2/2}e^{-\omega^2\sigma_{\Delta\tau}^2/2}\times \\
&\quad\left\{\left(e^{\beta}\mathbf{l}_{x_{q,m}}^\intercal + e^{-\beta}\mathbf{r}_{x_{q,m}}^\intercal\right) Re\left[\mathbf{c}_{y_{q,m}}e^{-j\omega\tau}\right]\right. \\
&+ \left. Re\left[e^{-j\omega\tau}\mathbf{c}_{x_{q,m}}^\intercal\right]\left(e^{\beta}\mathbf{l}_{y_{q,m}} + e^{-\beta}\mathbf{r}_{y_{q,m}}\right)\right\} \\
&+ 2\left(Re\left[\mathbf{c}_{x_{q,m}}^H\mathbf{c}_{y_{q,m}}\right] + e^{-2\omega^2\sigma_{\Delta\tau}^2}Re\left[\mathbf{c}_{x_{q,m}}^\intercal\mathbf{c}_{y_{q,m}}e^{-j2\omega\tau}\right]\right),
\end{aligned} \tag{B.10}$$

where:

$$\mathbf{l}_{x_{q,m}} = [X_{q,m-N+1}^{(l)}, \ \ldots, \ X_{q,m}^{(l)}]^\intercal - \mathbf{1}\sum_{k=m-N+1}^{m}\frac{X_{q,k}^{(l)}}{N}, \tag{B.11}$$

$$\mathbf{r}_{x_{q,m}} = [X_{q,m-N+1}^{(r)}, \ \ldots, \ X_{q,m}^{(r)}]^\intercal - \mathbf{1}\sum_{k=m-N+1}^{m}\frac{X_{q,k}^{(r)}}{N}, \tag{B.12}$$

$$\mathbf{c}_{x_{q,m}} = [X_{q,m-N+1}^{(c)}, \ \ldots, \ X_{q,m}^{(c)}]^\intercal - \mathbf{1}\sum_{k=m-N+1}^{m}\frac{X_{q,k}^{(c)}}{N}, \tag{B.13}$$

$$\beta = \frac{\ln(10)}{20}\gamma, \qquad \sigma_{\Delta\beta}^2 = \left(\frac{\ln(10)}{20}\right)^2\sigma_{\Delta\gamma}^2, \tag{B.14}$$

and similarly for the noisy/processed signal.

An expression for $E_\Delta\left[||\mathbf{x}_{q,m} - \mu_{\mathbf{x}_{q,m}}||^2\right]$ may be obtained from (B.10) by replacing all instances of $y_{q,m}$ by $x_{q,m}$ and vice versa for $E_\Delta\left[||\mathbf{y}_{q,m} - \mu_{\mathbf{y}_{q,m}}||^2\right]$.

The final DBSTOI measure is obtained by estimating the correlation coefficients, $\hat{\rho}_{q,m}$, for all frames, m, and frequency bands, q, in the signal and

averaging across these [7]:

$$\text{DBSTOI} = \frac{1}{QM} \sum_{q=1}^{Q} \sum_{m=1}^{M} \hat{\rho}_{q,m}, \tag{B.15}$$

where Q and M is the number of frequency bands and the number of frames, respectively.

It can be shown that whenever the left and right ear inputs are identical, the DBSTOI measure produces scores which are identical those of the monaural STOI (that is, the modified monaural STOI measure based on (B.5) and without clipping).

2.4 Determination of γ and τ

Finally, we consider the parameters γ and τ. These parameters are determined individually for each time unit, m, and third octave band, q, such as to maximize the final DBSTOI measure. Thus, each correlation coefficient estimate is a function of its own set of parameters, $\hat{\rho}_{q,m}(\gamma, \tau)$. The DBSTOI measure, (B.15), can therefore be maximized by maximizing each of the estimated correlation coefficients individually:

$$\hat{\rho}_{q,m} = \max_{\gamma,\tau} \quad \hat{\rho}_{q,m}(\gamma, \tau). \tag{B.16}$$

In practice, $\hat{\rho}_{q,m}$ is evaluated for a discrete set of γ and τ values and the highest value is chosen.

3 Results

The DBSTOI measure accepts binaural input signals which have been non-linearly processed, and is therefore applicable to a large range of acoustical conditions. We investigate the prediction performance of the measure in a selection of conditions: 1) speech masked by a single additive point noise source in the horizontal plane (a condition often used for evaluation of binaural intelligibility predictors [11, 12, 14–16]), 2) speech masked by a single point source with separate non-linear enhancement for each ear (to the knowledge of the authors, no other method is capable of providing predictions under such a condition), and 3) speech masked by multiple interferers and linearly processed with a beamformer.

3.1 Frontal Target and Point Interferer With and Without ITFS

First, we investigate the ability of the DBSTOI measure to predict SRTs[3]. Predictions are compared to the results of two experiments where SRTs were measured in normal hearing Danish subjects. In both experiments, we simulated a binaural anechoic environment by use of Head Related Transfer Functions (HRTFs) [24] and presented the resulting binaural signals through Sennheiser HDA200 headphones at a comfortable level. The target signal consisted of sentences from the Dantale II corpus [25], always emanating from a point source directly in front of the subject. The target signal was masked by a single point noise source, located in the horizontal plane. The further specifics of the two experiments are given by:

Experiment 1: Speech reception was measured in 10 conditions, differing only by the location of a single Speech Shaped Noise (SSN) interferer in the horizontal plane. The subjects listened to one five-word sentence at a time, repeating whatever words were heard. The experimenter marked the correctly identified words. The experiment was carried out for 10 normal hearing adult subjects. For each condition three repeated measurements were taken at 6 different Signal to Noise Ratios (SNRs). This resulted in the scoring of 10 subjects $\times$ 10 conditions $\times$ 6 SNRs $\times$ 3 repetitions $=$ 1800 sentences. ■

Experiment 2: Speech reception was measured in 9 conditions. Conditions 1–3 used SSN interferers at different positions in the horizontal plane. The left and right ear signals were independently subjected to ITFS with an Ideal Binary Mask (IBM) [26] as follows. The target and interferer signals were TF-decomposed with a short-time DFT, and TF-units with an SNR of less than 0 dB were attenuated by 10 dB. This finite attenuation was chosen to limit the improvement in intelligibility. Conditions 4–6 used the same interferer positions, but used bottle factory hall noise [27], rather than SSN. This noise type, which is a recording of bottles on a conveyor belt, has more energy at higher frequencies compared to SSN and is highly non-stationary. Conditions 4–6 did not include ITFS. Conditions 7–9 were the same as conditions 4–6 but with ITFS. The subjects were asked to select the words they heard on a screen. For each of the five words the subjects were shown 10 possible words, as well a pass-button to indicate that the given word had not been heard. A similar procedure is investigated in [28] and is shown to give results almost identical to those of the procedure used in experiment 1. Each condition was tested for 13 subjects, each at 6 SNRs and repeated 3 times. This resulted in the scoring of 13 subjects $\times$ 9 conditions $\times$ 6 SNRs $\times$ 3 repetitions $=$ 2106 sentences. ■

SRTs were determined individually for each subject for each condition

[3]The 50% SRT is the SNR where the subject scores 50% correct words.

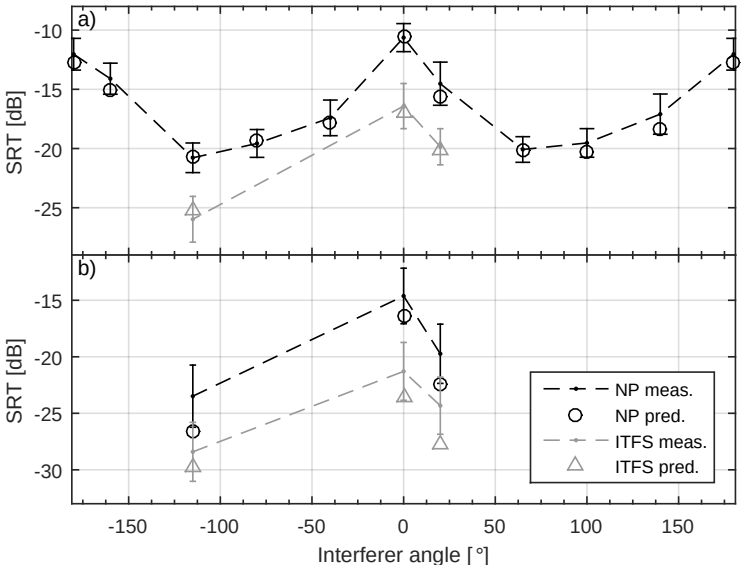


Fig. B.2: Measured (”meas.”) and predicted (”pred.”) SRTs with (”ITFS”) and without ITFS (”NP”). a) Frontal speech masked by a single SSN interferer and b) frontal speech masked by a single bottling factory hall noise interferer. The error bars indicate the standard deviation of the measured SRTs across subjects.

above. This was done by performing a maximum likelihood fit of a logistic function to the measured data as described in [29]. The SRTs were averaged across all subjects to obtain one mean SRT for each condition. To predict SRTs with the DBSTOI measure, a calibration constant was determined by scoring the condition with both target and SSN interferer in front, at an SNR equal to the SRT measured for this condition. SRT predictions in all the conditions were then made by adaptively varying the input SNR until the DBSTOI score was close to the calibration score.

The results of both measurements and predictions are shown in Fig. B.2. The upper plot, showing conditions with SSN masking, indicates that by calibrating the DBSTOI to a single condition, it is possible to predict the results of all the other combinations of ITFS and interferer positions to within less than one standard deviation of the measurements. In the bottom plot, showing conditions with bottle factory masking, a downward bias of 1-3 dB is seen in the predictions. This is most likely due to the method being calibrated to SSN masking. Note, though, that the relative effects of interferer position and ITFS are still predicted accurately.

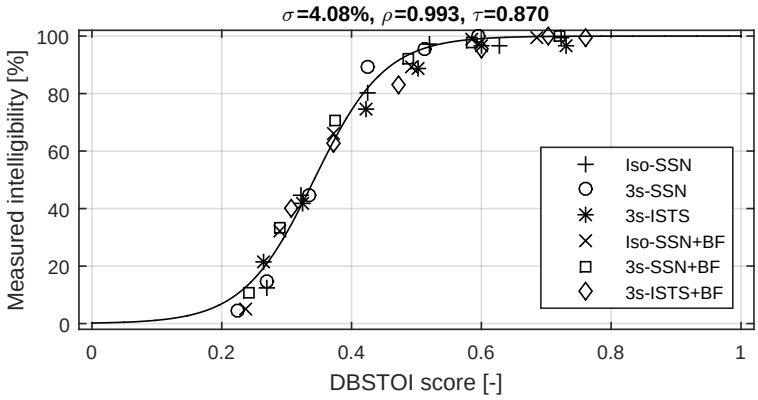


Fig. B.3: The results of the multiple-interferer experiment vs. computed DBSTOI scores. A fitted logistic function is shown too. Standard deviation (σ) and Pearson correlation (ρ) are computed relative to the fitted logistic function. The Kendall rank correlation (τ) is also shown.

3.2 Frontal Target and Multiple Interferers With/Without Beamforming

In this section, we evaluate the DBSTOI measure for a range of conditions with multiple interferers. An experiment similar to experiment 1 discussed in Section 3.1 was carried out. Ten normal hearing subjects were presented with Dantale II sentences in six different conditions for a scoring of 10 subjects $\times$ 6 conditions $\times$ 6 SNRs $\times$ 3 repetitions $=$ 1080 sentences in total. These were averaged across subjects and repetitions to produce a total of 36 data points. All six conditions were anechoic with speech originating from the front. The first condition was contaminated by isotropic (”Iso”) SSN. The second condition was contaminated by uncorrelated SSN from point sources at 110°, 180° and -110° in the horizontal plane (”3s”). The third condition considered the same layout of noise sources as condition two, but used three different segments of the International Speech Test Signal (ISTS) as noise [30]. Conditions four to six were the same as one to three, but include monaural 2-microphone Minimum Variance Distortionless Response (MVDR) beamforming as used in a behind-the-ear hearing aid (”BF”). DBSTOI scorings were made for each of the 6 conditions and 6 SNRs. Fig. B.3 shows the results. Although the investigated conditions are highly diverse, the DBSTOI predictions appear to be very well in line with the measured intelligibility.

3.3 Computational Cost

A key motivation for the DBSTOI measure is to avoid the necessity of Monte Carlo simulation, as was the case for the BSTOI measure proposed in [19]. It is therefore expected that the DBSTOI measure is computationally less de-

STOI	BSTOI	DBSTOI
5.3 s	1086.3 s	62.2 s

Table B.1: Time spent producing a scoring of 100 seconds of white noise on a Lenovo W530 with an Intel Core i7-3820QM, 2.7 GHz. The authors' own MATLAB implementations of the BSTOI and DBSTOI measures were used, while a STOI measure implementation was provided by the authors of [7].

manding than the BSTOI measure. To verify this, the measures were each used to score 100 seconds of white noise (the computational demand of computing the measures is independent of signal type). This was done simply with the `timeit`-function in MATLAB. The results are shown in Table B.1. Evaluating the DBSTOI measure is approximately 12 times more time consuming than evaluating the monaural STOI measure. Evaluating the Monte-Carlo-based BSTOI measure is, however, more than 17 times as time consuming as evaluating the proposed DBSTOI measure.

4 Conclusions

In this paper we present and investigate a binaural speech intelligibility measure, DBSTOI, which accepts input signals that have been processed by e.g. a speech enhancement algorithm. The measure is obtained by combining the Short-Time Objective Intelligibility (STOI) measure with a modified Equalization Cancellation (EC) stage. The presented measure improves upon the previously proposed BSTOI measure by providing deterministic results at a lower computational cost. We demonstrate that the measure is able to predict accurately the effect on intelligibility of simultaneous non-linear signal enhancement and binaural advantage, in addition to simpler conditions with non-linear enhancement or binaural advantage separately. We furthermore show that the computational costs associated with the proposed DBSTOI measure is more than 17 times lower than those of the previously proposed BSTOI measure.

References

[1] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[2] American National Standards Institute, "S3.5-1997: Methods for calculation of the speech intelligibility index," 1997.

[3] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Matching pursuit for channel selection in coclear implants based on an intelligibility metric,"

in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, Aug. 2012, pp. 504–508.

[4] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[5] T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355–367, Jan. 1971.

[6] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[7] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for inteligibility prediction of time-frequency weighted noisy speech," *IEEE Tran. on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.

[8] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE Tran. on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 430–440, Feb. 2014.

[9] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3679–3689, Dec. 2004.

[10] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, July 2014.

[11] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 331–342, Apr. 2006.

[12] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension and evaluation of a binaural speech intelligibility model," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2479–2497, Dec. 2010.

[13] J. Rennies, T. Brand, and B. Kollmeier, "Prediction of the intelligibility of reverberation on binaural speech intelligibility in noise and in quiet," *J. Acoust. Soc. Am.*, vol. 130, no. 5, pp. 2999–3012, Nov. 2011.

[14] M. Lavandier and J. F. Culling, "Prediction of binaural speech intelligibility against noise in rooms," *J. Acoust. Soc. Am.*, vol. 127, no. 1, pp. 387–399, Jan. 2010.

[15] S. Jelfs, J. F. Culling, and M. Lavandier, "Revision and validation of a binaural model for speech intelligibility in noise," *Hearing Research*, vol. 275, no. 1–2, pp. 96–104, May 2011.

[16] M. Lavandier, S. Jelfs, J. F. Culling, A. J. Watkins, A. P. Raimond, and S. J. Makin, "Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 218–231, Jan. 2012.

[17] R. Wan, N. I. Durlach, and H. S. Colburn, "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," *J. Acoust. Soc. Am.*, vol. 128, no. 6, pp. 3678–3690, Sept. 2010.

[18] A. W. Bronkhorst, "The cocktail party phenomenon: A review on speech intelligibility in multiple-talker conditions," *Acta Acustica United with Acustica*, vol. 86, no. 1, pp. 117–128, Jan. 2000.

[19] A. H. Andersen, J. M. de Haan, Z. Tan, and J. Jensen, "A binaural short time objective intelligibility measure for noisy and enhanced speech," in *INTERSPEECH*, Dresden, Germany, Sept. 2015, pp. 2563–2567.

[20] N. I. Durlach, "Equalization and cancellation theory of binaural masking-level differences," *J. Acoust. Soc. Am.*, vol. 35, no. 8, pp. 1206–1218, Aug. 1963.

[21] N. I. Durlach, "Binaural signal detection: Equalization and cancellation theory," in *Foundations of Modern Auditory Theory Volume II*, Jerry V. Tobias, Ed., pp. 371–462. Academic Press, New York, 1972.

[22] H. vom Hövel, *Zur bedeutung der übertragungseigenschaften des aussenohrs sowie des binauralen hörsystems bei gestörter sprachübertragung*, Ph.D. thesis, Rheinisch-Westfälische Technische Hochschule Aachen, 1984.

[23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech," *J. Acoust. Soc. Am.*, vol. 130, no. 5, pp. 3013–3027, Nov. 2011.

[24] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proccedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct. 2001.

[25] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *International Journal of Audiology*, vol. 42, no. 1, pp. 10–17, Jan. 2003.

[26] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 123, no. 3, pp. 1673–1682, Mar. 2008.

[27] M. Vestergaard, "The eriksholm cd 01: Speech signals in various acoustical environments," Tech. Rep. 050-08-01, Oticon Research Centre, Eriksholm, Snekkersten, 1998.

[28] E. R. Pedersen and P. M. Juhl, "User-operated speech in noise test: Implementation and comparison with a traditional test," *International Journal of Audiology*, vol. 53, no. 5, pp. 336–344, May 2014.

[29] T. Brand and B. Kollmeier, "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," *J. Acoust. Soc. Am.*, vol. 111, no. 6, pp. 2801–2810, June 2002.

[30] I. Holube, S. Fredelake, M. Vlaming, and B. Kollmeier, "Development and analysis of an international speech test signal (ISTS)," *International Journal of Audiology*, vol. 49, no. 12, pp. 891–903, Dec. 2010.

Paper C

Speech Intelligibility Prediction as a Classification Problem

Asger Heidemann Andersen
Esther Schoenmaker
Steven van de Par

The paper has been presented at the
26th IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-6, Salerno, Italy, 2016.

© 2016 IEEE
The layout has been revised.

Abstract

Speech Intelligibility Prediction (SIP) algorithms are becoming increasingly popular for objective evaluation of speech processing algorithms and transmission systems. Most often, SIP algorithms aim to predict the average intelligibility of an average listener in some specific listening condition. In the present work, we instead consider the aim of predicting the intelligibility of single words. I.e. we attempt to predict whether or not a subject in a listening experiment was able to correctly repeat a particular word. We base the prediction on a noisy and potentially processed/degraded recording of the spoken word (as presented to a subject), as well as a clean reference recording of the spoken word. The problem can be treated as a supervised binary classification problem of predicting whether a specific word will or will not be understood. We investigate a number of different ways to extract features from the degraded and clean speech samples. The classification is carried out by means of Fisher discriminant analysis. Despite the large variability of speech intelligibility experiments, it is possible to obtain a considerable degree of predictive power.

1 Introduction

Speech intelligibility is an important measure of performance for systems and algorithms concerned with storage, transmission, processing and reproduction of speech. However, measuring speech intelligibility is a time consuming, and thus expensive, task involving many human subjects. Therefore, computational predictors of intelligibility are gaining popularity for evaluation of algorithms and systems involving speech. Such predictors are typically based on mathematical models of how the human ear and brain extracts information from complex sound signals.

The task of Speech Intelligibility Prediction (SIP) was first studied as a tool for improving the intelligibility of telephone systems [1]. This resulted in the development of the Articulation Index (AI) which has later been refined and standardized as the Speech Intelligibility Index (SII) [2]. The SII provides an *index* (an objective rating of intelligibility) in the range 0 to 1, based on information about the long-term spectra of speech and noise. The relationship between this index and actual intelligibility is dependent on e.g. the speech material. The SII is designed for predicting intelligibility in conditions where speech and additive stationary noise is presented to one ear only. A number of variations of the SII has been introduced with the aim to extend the range of conditions to which it is applicable. The Extended SII (ESII) computes the SII in short time frames and thereby allows for handling fluctuating noise [3, 4]. The Coherence SII (CSII) predicts intelligibility of speech which has been non-linearly degraded [5]. The Binaural Speech Intelligibility Measure (BSIM) extends the SII with an Equalization Cancellation (EC) stage [6]

such as to predict intelligibility in conditions with binaural advantage [7].

Recently, the Short-Time Objective Intelligibility (STOI) measure [8] has gained popularity for evaluation of signal processing algorithms and systems. The STOI measure predicts intelligibility from the correlation between third-octave envelopes of the degraded signal and a clean reference signal, and thus does not require knowledge of the noise in separation. The STOI measure has shown to compare favorably with other measures of intelligibility, for processed and degraded speech [9–11].

The above mentioned methods are designed for predicting the average intelligibility in a given acoustic condition, with a given noise type and/or with a specific type of speech processing or degradation. This prediction is typically carried out on the basis of a substantial amount of speech (i.e. at least several sentences). A less studied parallel problem is that of microscopic SIP, which aims to predict the intelligibility of individual speech tokens such as words or logatomes. Microscopic SIP may be useful as a tool for fine tuning of speech processing algorithms, and can provide insights into how humans understand speech. Microscopic SIP has been studied in a number of different settings by use of Automatic Speech Recognition (ASR) systems [12–14]. In [15] it is hypothesized that intelligibility is related to the availability of spectro-temporal regions in which the target speech is dominating: *glimpses*. It is shown that the quantity and quality of such glimpses correlate well with measured intelligibility.

The described intelligibility predictors are mainly based on different combinations of auditory models and heuristics. However, the microscopic task of predicting whether or not a particular speech token was heard correctly by a specific subject can be considered as a binary classification problem. In the present study we extract different types of features from clean and noisy speech tokens. The extracted feature vectors are used together with Fisher discriminant analysis to produce specific predictions of whether or not the token was heard correctly by a subject in a listening experiment. As training and testing data, we use the results of a listening experiment described in [16, 17]. To the knowledge of the authors, such an approach has not previously been applied to the microscopic SIP problem.

The remainder of the paper progresses as follows. In Section 2, the used speech intelligibility database is briefly presented. In Section 3, we describe how features are extracted from clean and noisy speech tokens. Section 4 describes how Fisher discriminant analysis is used to provide specific predictions of intelligibility for each speech token. Section 5 presents and discusses results. Section 6 concludes upon the presented findings.

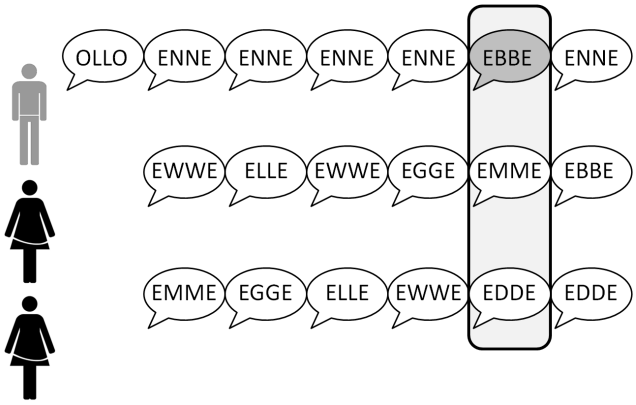


Fig. C.1: An illustration of the logatome experiment.

2 Experimental Data

For training and testing of the discussed classifier, we use results from a listening experiment presented in [16, 17]. The experiment measured intelligibility in four different conditions. The purpose of the experiment was to investigate what binaural factors influence speech intelligibility. It has been found that when the speaker of interest is spatially separated from the interfering sound sources, an improvement in speech intelligibility can be found, as compared to having all sources at the same location [18]. The study investigated specifically whether the instantaneous difference in spatial cues for the target speaker and interferers is important for this improved intelligibility. This was accomplished by using a stimulus manipulation, termed Equal Local Azimuth (ELA), that removes these differences in spatial cues by locally (in time and frequency) presenting all sources from one common azimuthal direction while conserving the overall spatial image [16, 17].

The experiment was based on the Oldenburg Logatome Corpus [19]. In this, different sequences of six logatomes are spoken simultaneously by three different speakers as illustrated in Fig. C.1. The target speaker is indicated by one heading logatome placed before the onset of the interfering speakers. The objective of the subject is to repeat the deviating logatome of the target speaker, as marked in gray on Fig. C.1, by identifying it in a list of 7 possibilities (forced choice). The deviating logatome is placed randomly on one of the three last positions.

The target and interfering speakers were placed at different angles in the frontal horizontal plane in an anechoic environment, by application of Head Related Transfer Functions (HRTFs). Two different types of spatial layouts were used in the experiment: 1) the three speakers were collocated on a uni-

Condition	Spatial layout	Processing
1	Collocated	None
2	Separated	None
3	Collocated	ELA
4	Separated	ELA

Table C.1: An overview of the four conditions in the experiment.

formly randomly chosen position in the frontal horizontal plane, and 2) the three speakers were located at different uniformly randomly chosen positions in the frontal horizontal plane. In the latter, positions were chosen such that any two speakers are always separated by at least 10°.

The two different spatial conditions combined with either the absence or presence of ELA processing leads to four different conditions which are summarized in Table C.1.

While only four different conditions were included in the experiment, it is worthwhile to consider the great variability inherent in the conditions. Especially the *separated* conditions vary greatly in difficulty depending on the randomly chosen positions of the speakers. This is mainly due to the fact that increased spatial separation between target and interfering speakers generally lead to improved intelligibility [18].

The experiment was carried out with 18 normal-hearing subjects. Each subject was presented with (4 conditions) $\times$ (168 trials/condition) $=$ 672 trials. This resulted in the collection of results for a total of (672 trials/subject) $\times$ (18 subjects) = 12096 trials. The overall results for the four conditions are shown in Fig. C.2. Most notably, intelligibility is significantly improved in the conditions with spatial separation. Further, intelligibility in the *separated* condition is decreased by the addition of ELA processing. See [17] for a detailed discussion of these results, and the impact of ELA processing.

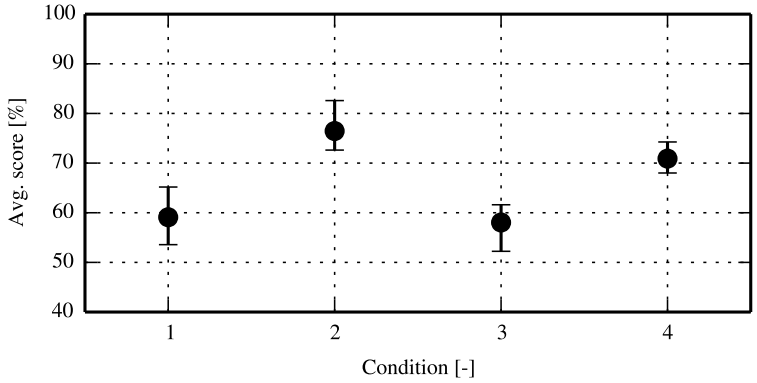


Fig. C.2: Mean subject score in the four conditions, numbered according to Table C.1. The error bars indicate the 25th and 75th percentiles.

3 Feature Extraction

The aim of this study is to predict, for each individual presented stimulus, whether or not the subject was able to correctly repeat the deviating logatome token of the target speaker. We therefore base predictions on only the part of the signal in which this deviating logatome is pronounced. We assume access to the speech mixture as presented to the left and right ears of the subject, as well as a clean reference signal for the left and right ears of the subject (i.e. the target speaker in the absence of interfering speakers). We investigate three different types of features which are all based on the correlation between the clean and noisy signals.

Instead of extracting features separately from the left and right ear signals, we make assumptions on how listeners combine information from the left and right ears. A highly successful model of this process assumes that the ears are used in a way similar to a suboptimal two-microphone adaptive beamformer [6]. In the pursuit of simplicity, however, we simply assume that the listener always attends to the ear that contributes most positively to intelligibility.

In the remainder of this section, we present the three investigated feature types.

3.1 STOI-Type Features

The STOI measure has shown very successful at predicting the intelligibility of noisy speech which has been processed by different speech enhancement algorithms [8]. We therefore include a type of features which is strongly

	# window lengths, L	Total features
STOI-1	1	15
STOI-2	2	30
STOI-4	4	60
STOI-8	8	120
PEMO-1	1	33
PEMO-2	2	66
PEMO-4	4	132
PEMO-8	8	264
MFCC-1	1	26
MFCC-2	2	52
MFCC-4	4	104
MFCC-8	8	208

Table C.2: An overview of the investigated types of features.

inspired by the STOI measure.

The clean and noisy audio signals are first resampled to 10 kHz and represented in the short-time frequency domain by use of the Discrete Fourier Transformation (DFT). This is done exactly as for the STOI measure, and with the same choice of parameters [8]. We denote the k'th frequency bin of the m'th frame of the left clean signal as $\hat{x}_{k,m}^{(l)}$. Similarly, the corresponding DFT coefficient of the right clean signal and the left and right degraded signals are denoted by $\hat{x}_{k,m}^{(r)}$, $\hat{y}_{k,m}^{(l)}$ and $\hat{y}_{k,m}^{(r)}$, respectively.

Envelopes are extracted in $Q = 15$ third octave bands:

$$X_{q,m}^{(l)} = \sqrt{\sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m}^{(l)}|^2}, \tag{C.1}$$

where $k_1(q)$ and $k_2(q)$ are the lower and upper limits, respectively, of the q'th third octave band. Similarly defined are envelope samples for the other three signals: $X_{q,m}^{(r)}$, $Y_{q,m}^{(l)}$ and $Y_{q,m}^{(r)}$. These envelopes are arranged in short frames of N samples:

$$\mathbf{x}_{q,m|N}^{(l)} = [X_{q,m-N+1}^{(l)}, \ldots, X_{q,m}^{(l)}]^\intercal. \tag{C.2}$$

Similar frames $\mathbf{x}_{q,m|N}^{(r)}$, $\mathbf{y}_{q,m|N}^{(l)}$ and $\mathbf{y}_{q,m|N}^{(r)}$, are defined for the three other signals. Note that the frames are generated with an overlap of $N-1$ samples. The normalized correlation coefficient between clean and degraded

envelopes is then computed:

$$\rho_{q,m|N}^{(l)} = \frac{(\mathbf{x}_{q,m|N}^{(l)} - \mathbf{1}\mu_{\mathbf{x}_{q,m|N}^{(l)}})^{\mathsf{T}}(\mathbf{y}_{q,m|N}^{(l)} - \mathbf{1}\mu_{\mathbf{y}_{q,m|N}^{(l)}})}{||\mathbf{x}_{q,m|N}^{(l)} - \mathbf{1}\mu_{\mathbf{x}_{q,m|N}^{(l)}}|| \cdot ||\mathbf{y}_{q,m|N}^{(l)} - \mathbf{1}\mu_{\mathbf{y}_{q,m|N}^{(l)}}||}, \tag{C.3}$$

where $\mu_{(\cdot)}$ is the mean of the entires in the respective vector, $\mathbf{1}$ is a vector of ones, and $||\cdot||$ is the Euclidean norm. A similar correlation coefficient, $\rho_{q,m|N}^{(r)}$, is defined for the right ear signals. The final features are computed as the better-ear correlation averaged across the length of one token:

$$s_{q|N} = \sum_{m=N}^{M} \max\left(\rho_{q,m|N}^{(l)}, \rho_{q,m|N}^{(r)}\right), \tag{C.4}$$

where M is the number of samples in the computed envelopes. This extracts $Q = 15$ features from one speech token. We may compute the features for multiple window lengths, $N_1, \ldots, N_L$, and stack them into one feature-vector:

$$s = \left[s_{1|N_1}, \ldots, s_{Q|N_1}, s_{1|N_2}, \ldots, s_{Q|N_L}\right]^{\mathsf{T}}. \tag{C.5}$$

We evaluate the use of 1, 2, 4 and 8 window lengths ($L = 1, 2, 4, 8$) spaced logarithmically between 30 ms and 1 s (in the $L = 1$ case, the window length is set to the geometric mean of 30 ms and 1 s). The envelope-representation of the signals, (C.1), is zero-padded such as to ensure that at least one envelope-vector, (C.2), can be formed.

3.2 Features Based on an Auditory Model

The STOI measure has shown to be very successful in many respects. However, it uses a very simple DFT-based filterbank to model the human auditory system. To investigate whether this adversely affects prediction performance, we include a type of features which are very similar to those described in Section 3.1 but which rely on a more sophisticated auditory model based on [20]. We refer to these as PEMO-features after the auditory model on which they are based. The model is composed of:

1. A 33 band gammatone filterbank with center frequencies logarithmically spaced from 80 to 8000 Hz.
2. Envelope extraction by single rectification and lowpass filtering [20].
3. Adaptive compression loops [20].

The model was implemented by use of the Auditory Modeling Toolbox [21].

The output of the auditory model is 33 signals with a sampling rate equal to that of the input signal. We generate such internal representations for both the clean signal and the degraded signal, and derive features in the same way as was done for the envelopes described in Section 3.1. Namely, for each frequency band, the internal representations are segmented into short-time frames, as described by (C.2), the correlation between clean and degraded frames is computed, as described by (C.3), and the correlation coefficients are averaged across time in a better-ear fashion, as described by (C.4). Frames are generated with an overlap of 7/8 the frame length (due to the higher sampling frequency, it is not realistic to generate windows with an increment of only one sample, as is done in Section 3.1). Feature vectors with multiple frame lengths are generated in the same manner as described in Section 3.1, using the same selections of window lengths.

3.3 Mel Frequency Cepstral Features

Mel Frequency Cepstral Coefficients (MFCCs) have shown to provide good performance in ASR systems. This could be interpreted as evidence that these capture much of the relevant information necessary for speech understanding, and thereby suggests that they could be useful for SIP. We therefore include a type of features based on the correlation between MFCCs of the clean and degraded signals.

MFCCs are extracted from the clean and degraded signals with the MATLAB tool made available by [22]. Default settings for extracting MFCCs were used. This returns 13 coefficients per 25-ms window. An additional 13 first order derivatives are added [22]. The MFCCs of the clean and degraded signals are arranged in short time frames and correlated as described in Section 3.1. Feature vectors with multiple frame lengths are generated in the same manner as described in Section 3.1, using the same selections of window lengths. An overview of all the investigated sets of features is provided in Table C.2.

4 Classification

Due to the preliminary nature of the present study and the large number of different feature types investigated, a very simple classification procedure is used[1]. The 12096 data points are randomly separated into training and test-

[1]Classification was carried out using tools from the software package `scikit-learn` [23]. Several different classification algorithms were tested with the available data. Using more sophisticated classifiers it was possible to increase performance to a small extent. However, these results were not found to add anything significant to the conclusions of this study, and are therefore not included.

ing datasets of equal size. The features are projected onto a single dimension by use of Fisher linear discriminant analysis [24]:

$$\tilde{z}_i = \mathbf{w}^\mathsf{T}\mathbf{z}_i, \qquad i = 1, \ldots, 12096, \tag{C.6}$$

where $\mathbf{z}_i$ is the feature vector of the i'th token, which is projected onto a one-dimensional representation, $\tilde{z}_i$, by use of the weighting vector, $\mathbf{w}$ [24]:

$$\mathbf{w} = (\mathbf{\Sigma}_0 + \mathbf{\Sigma}_1)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \tag{C.7}$$

where $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ are the covariance matrices of features in the training set for which the subjects answered, respectively, incorrectly and correctly, while $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ are the corresponding means. We refer to $\tilde{z}_i$ as a decision variable. From $\tilde{z}_i$, we use simple thresholding to arrive at a prediction, c_i , of whether or not the token was correctly answered ($c_i = 1$ corresponding to predicting a correct answer). This is done by maximum likelihood, assuming normally distributed decision variables:

$$c_i = \begin{cases} 1, & \text{if} \quad p_0\,\mathcal{N}(\tilde{z}_i; \tilde{\mu}_0, \tilde{\sigma}_0^2) < p_1\mathcal{N}(\tilde{z}_i; \tilde{\mu}_1, \tilde{\sigma}_1^2), \\ 0, & \text{otherwise}, \end{cases} \tag{C.8}$$

where $\tilde{\mu}_0$ and $\tilde{\mu}_1$ are the means of the decision variables of correctly and incorrectly answered tokens in the training dataset, $\tilde{\sigma}_0^2$ and $\tilde{\sigma}_1^2$ are the corresponding variances, p_0 and p_1 are prior probabilities of, respectively, $c_i = 0$ and $c_i = 1$ (computed from the training dataset), and $\mathcal{N}(x; \mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2.

The full dataset contains results of 12096 tokens, 66.1% of which have been answered correctly. Thereby, a naive classifier that predicts any token to have been correctly answered, can obtain an accuracy of 66.1%. To study the classification performance in the absence of this asymmetry, we generate a symmetric dataset containing all incorrectly answered tokens as well as an equally sized, randomly sampled, subset of the correctly answered tokens. We refer to this as the "50/50" dataset. The 50/50-dataset is also split into equally large training and testing datasets.

5 Results and Discussion

We show results for the 12 different sets of features (Table C.1), each for the full- and the 50/50-dataset. The performance is scored in terms of: 1) accuracy, 2) F_1-measure, and 3) Cohen's Kappa [25], which is a simple rescaling of accuracy that accounts for the agreement occurring by chance.

The performance measures are shown in Table C.3. Considering the full dataset, all types of features lead to similar accuracies, which are at most slightly better than what could have been obtained by simply classifying all

	Full data			50/50 data		
	A	F_1	κ	A	F_1	κ
STOI-1	0.656	0.779	0.002	0.592	0.634	0.182
STOI-2	0.665	0.782	0.028	0.630	0.665	0.257
STOI-4	0.677	0.787	0.064	0.642	0.658	0.281
STOI-8	0.685	0.787	0.084	0.637	0.649	0.272
PEMO-1	0.671	0.783	0.044	0.635	0.636	0.268
PEMO-2	0.688	0.789	0.094	0.653	0.657	0.304
PEMO-4	**0.691**	0.791	**0.103**	**0.663**	0.674	**0.324**
PEMO-8	0.690	0.787	0.100	0.653	0.663	0.304
MFCC-1	0.678	0.790	0.066	0.633	0.652	0.263
MFCC-2	0.681	0.789	0.075	0.637	0.660	0.271
MFCC-4	0.689	**0.792**	0.098	0.656	**0.676**	0.309
MFCC-8	0.684	0.786	0.082	0.651	0.672	0.299

Table C.3: An overview of the test performance obtained with Fisher linear discriminant analysis for the investigated sets of features. Performance is given in terms of prediction accuracy, A, F_1-measure, and Cohen's κ. Results are shown for the full dataset with 66.1% correct answers and the reduced dataset with 50% correct answers.

tokens as correctly answered (66.1%). This issue is further supported by the κ-measures which are only slightly larger than zero for all feature types. Due to the asymmetry of the data, it is, however, not possible to conclude from this that the trained classifiers provide no predictive power. This is seen by considering panel "a)" of Fig. C.3 which shows the distribution of decision variables for correctly and incorrectly answered tokens. It is clearly seen that correctly answered tokens tend to have higher decision variables. On the contrary, due to the excess of correctly answered tokens, the large majority of tokens are most likely to be answered correctly, based on the evidence of the decision variable.

The problem of asymmetry is alleviated by the 50/50-dataset, at the price of having a slightly smaller dataset of 8196 tokens. From Table C.3 it is seen that the reduced dataset leads to slightly lower accuracy. On the contrary, a naive classifier can only obtain an accuracy of 50% in this case; significantly lower than what is obtained for any of the feature types. This observation is reflected in the κ-measures which indicate prediction performance significantly greater than chance. A cause of this improvement in performance can be seen in panel "c)" of Fig. C.3. The separation between correctly and incorrectly answered tokens is similar to that of the full dataset, but due to the decreased prior probability of tokens being correctly answered, tokens with negative decision variables now have a large probability of being incorrectly answered.

The different feature types generally lead to similar performance. It can,

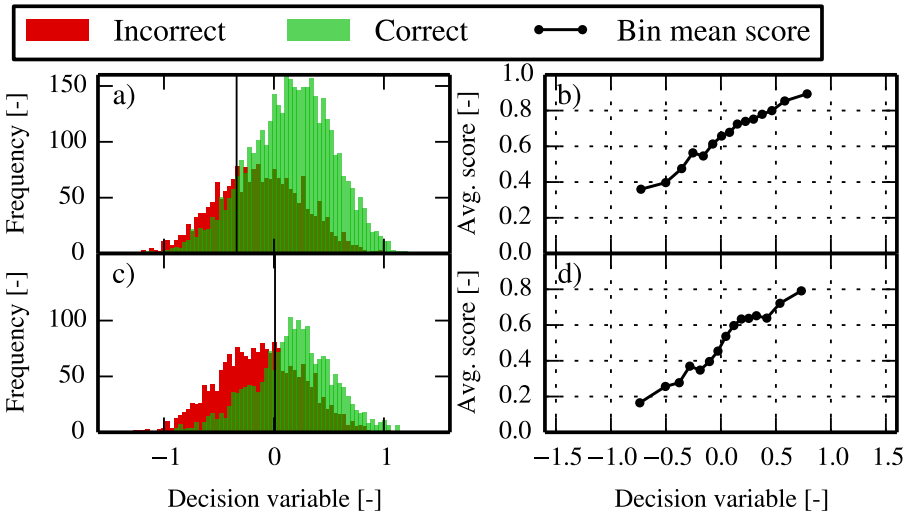


Fig. C.3: Detailed results for the PEMO-4 features which lead to the highest obtained accuracy. a) The distribution of decision variables of tokens which have been correctly and incorrectly answered in the full dataset. b) The tokens of the full dataset were sorted in 15 equally large bins according to decision variable. The plot shows the fraction of correctly answered tokens for each bin versus the median decision variable of the bin. c) The same as ”a)”, but for the 50/50-dataset. d) The same as ”b)” but for the 50/50 dataset.

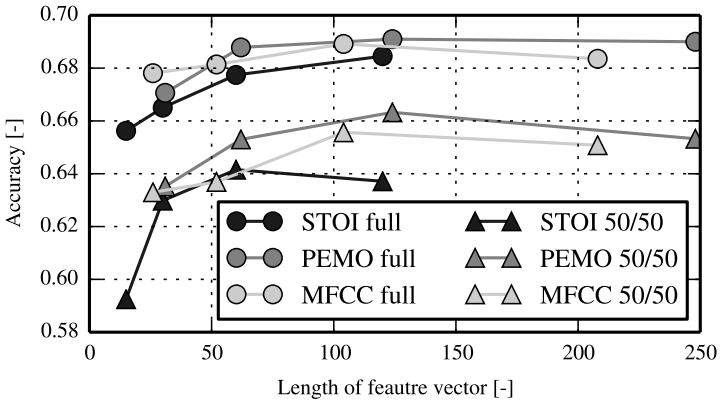


Fig. C.4: Prediction accuracy versus length of feature vector. For each feature type and dataset, the plot shows the accuracy for the four different values of L.

however, be observed that: 1) Increasing the feature dimensionality generally leads to increased performance, and 2) the PEMO and MFCC features give slightly higher performance than the STOI features. The second point can, however, simply be due to the fact that the PEMO and MFCC features have higher dimensionality than the STOI features (see Table C.2). The relationship between feature dimensionality and prediction accuracy is shown in Fig. C.4. This indicates that the PEMO and MFCC features perform slightly better than the STOI features, even for similar feature dimensionality. However, the largest impact appears to come from the feature dimensionality.

As shown in Fig. C.2, the performance of the subjects varied significantly between the four investigated conditions. These results are compared to predictions with the PEMO-4 features in Fig. C.5. This indicates that the predictions accurately capture the differences between the four conditions, but that a significant upwards bias is present. This is a natural consequence of the asymmetry of the dataset (i.e. in most cases it is best to predict a token to be correctly answered because of the high prior probability of this being the case) and is a good illustration of the differences between the objective of microscopic and macroscopic intelligibility prediction. It is noteworthy that the classifier correctly predicts a large advantage in the *separated* conditions (2 and 4) in spite of the assumption that listeners rely only on listening to the better ear. It is unknown whether prediction performance could be increased by using a more sophisticated model of how listeners combine information from the left and right ears (e.g. [6]).

The presented results may appear somewhat discouraging. However, the obtained accuracy should be viewed in the light of the fact that the stimuli is only one of many sources of variance in a SIP experiment. Fig. C.6 shows the measured and predicted performance of the 18 subjects individually. Also here, a large upwards bias is seen. Furthermore, it is evident that there are big differences between the performance of the individual subjects. These differences are not apparent in the predictions. This is not surprising as the differences most likely stem from differing levels of hearing and skill in the subjects: information which is not available in the speech stimuli. This is a rather large source of error, which no classifier can account for (unless given explicit information about subjects). Also worth mentioning, is the fact that the experiment was carried out with a 7-alternative forced choice procedure (Section 2). I.e. whenever a subject was unable to provide a motivated answer, there was a $1/7 = 14.3\%$ chance that the subject randomly picked the correct answer and the token was marked as correctly answered. In total, 33.9% of the tokens were incorrectly answered. Assuming that all incorrect answers were completely unmotivated, one should expect that $33.9\% \cdot 1/6 = 5.65\%$ of the tokens are "falsely" marked as correctly answered, because the subject guessed the correct answer by chance. A good classifier can predict that a token is not intelligible to an average subject, but has no chance of predicting

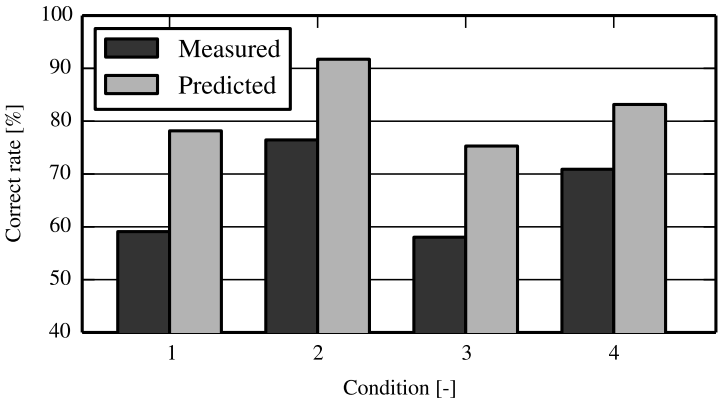


Fig. C.5: Predictions for the PEMO-4 features compared with the measured scores for the 4 conditions in the full dataset. The predictions are simply computed as the fraction of tokens, in each condition, predicted to be answered correctly.

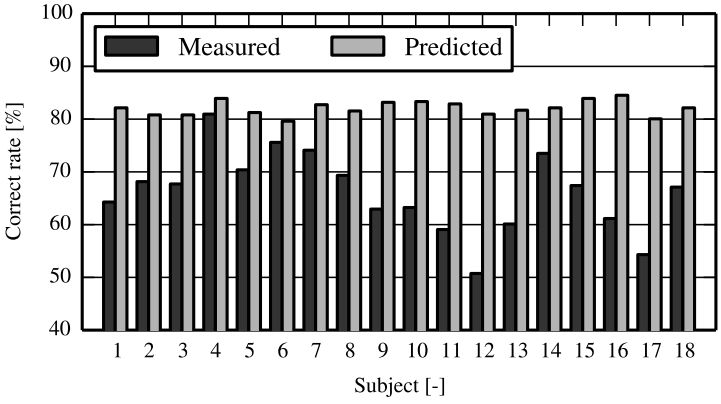


Fig. C.6: Predictions for the PEMO-4 features compared with the measured scores for the 18 subjects in the full dataset. The predictions are simply computed as the fraction of tokens, for each subject, predicted to be answered correctly.

whether the correct answer was randomly guessed. Even further, psychological aspects such as attention span, concentration and fatigue most certainly also have an impact on such an experiment. Such things cannot be accounted for by a classifier either. In general, one should expect that the results of such an experiment could vary significantly, on the microscopic level, if repeated with exactly the same subjects, noise realizations, etc. Such a repeated experiment could be used as a predictor for the original experiment. The accuracy of such a prediction could be considered as an upper bound for the performance obtainable by a classifier (and thereby also as a fair frame of reference to which performance can be compared).

6 Conclusions

We have investigated the task of microscopic intelligibility prediction as a classification problem. This was done by extracting features from individual speech-masked logatome tokens and using a Fisher linear discriminant classifier to predict whether a subject in a listening experiment could correctly repeat the token. Three different types of correlation-based features were investigated: 1) features similar to those used in the short-term objective intelligibility (STOI) measure, 2) features based on a more sophisticated auditory model, 3) features based on Mel frequency cepstral coefficients (MFCCs). All feature types resulted in significant predictive power. However, given the large inherent variability of listening experiments, it is uncertain what prediction accuracy is practically attainable. Performance differences appeared to stem mainly from differences in dimensionality between the different feature types. As a direction of further study, we propose repeating the same realization of a listening experiment multiple times, such as to study how much variability is strictly stochastic or due to human performance, and is thus not related to the specific speech tokens. This could be used to derive an upper bound of possible classification performance.

References

[1] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[2] ANSI, "Methods for calculation of the speech intelligibility index," *S3.5-1997*, 1997.

[3] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.

[4] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, Dec. 2006.

[5] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, Apr. 2005.

[6] N. I. Durlach, "Equalization and cancellation theory of binaural masking-level differences," *J. Acoust. Soc. Am.*, vol. 35, no. 8, pp. 1206–1218, Aug. 1963.

[7] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension and evaluation of a binaural speech intelligibility model," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2479–2497, Dec. 2010.

[8] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for inteligibility prediction of time-frequency weighted noisy speech," *IEEE Tran. on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.

[9] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[10] S. Jørgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acta Acustica United with Acustica*, vol. 101, pp. 1016–1025, 2015.

[11] K. Smeds, A. Leijon, F. Wolters, A. Hammarstedt, S. Båsjö, and S. Hertzman, "Comparison of predictive measures of speech recognition after noise reduction processing," *J. Acoust. Soc. Am.*, vol. 136, no. 3, pp. 1363—-1374, Sept. 2014.

[12] I. Holube and B. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *J. Acoust. Soc. Am.*, vol. 100, no. 3, pp. 1703–1716, 1996.

[13] M. R. Schädler, A. Warzybok, S. Hochmuth, and B. Kollmeier, "Matrix sentence intelligibility prediction using an automatic speech recognition system," *International Journal of Audiology*, vol. 54, pp. 1–8, 2015.

[14] M. Geravanchizadeh and A. Fallah, "Microscopic prediction of speech intelligibility in spatially distributed speech-shaped noise for normal hearing listeners," *J. Acoust. Soc. Am.*, vol. 138, no. 6, pp. 4004–4015, Dec. 2015.

[15] M. Cooke, "Glimpsing model of speech perception," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, Mar. 2006.

[16] E. Schoenmaker and S. van de Par, "Auditory streaming in cocktail parties and the extent of binaural benefit," in *The 21st International Congress on Acoustics*, Montreal, Canada, Jan. 2013, Acoustical Society of America.

[17] E. Schoenmaker, T. Brand, and S. van de Par, "The multiple contributions of interaural differences to improved speech intelligibility in multitalker scenarios," *J. Acoust. Soc. Am.*, vol. 139, no. 5, pp. 2589—-2603, May 2016.

[18] A. W. Bronkhorst, "The cocktail party phenomenon: A review on speech intelligibility in multiple-talker conditions," *Acta Acustica United with Acustica*, vol. 86, no. 1, pp. 117–128, Jan. 2000.

[19] B. T. Meyer, T. Jürgens, T. Wesker, T. Brand, and B. Kollmeier, "Human phoneme recognition depending on speech-intrinsic variability," *J. Acoust. Soc. Am.*, vol. 128, no. 5, pp. 3126—-3141, 2010.

[20] T. Dau, D. Püschel, and A. Kohlrausch, "A Quantitative Model of the "Effective" Signal Processing in the Auditory System. I. Model Structure," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3615–3622, June 1996.

[21] P. Søndergaard and P. Majdak, "The auditory modeling toolbox," in *The Technology of Binaural Listening*, Jens Blauert, Ed., pp. 33–56. Springer, Berlin, Heidelberg, 2013.

[22] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[25] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

Paper D

Predicting the Intelligibility of Noisy and Non-Linearly Processed Binaural Speech

Asger Heidemann Andersen
Jan Mark de Haan
Zheng-Hua Tan
Jesper Jensen

The paper has been published in the
IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 24, No. 11, pp. 1908–1920, 2016.

© 2016 IEEE
The layout has been revised.

Abstract

Objective speech intelligibility measures are gaining popularity in the development of speech enhancement algorithms and speech processing devices such as hearing aids. Such devices may process the input signals non-linearly and modify the binaural cues presented to the user. We propose a method for predicting the intelligibility of noisy and non-linearly processed binaural speech. This prediction is based on the noisy and processed signal as well as a clean speech reference signal. The method is obtained by extending a modified version of the short-time objective intelligibility (STOI) measure with a modified equalization-cancellation (EC) stage. We evaluate the performance of the method by comparing the predictions with measured intelligibility from four listening experiments. These comparisons indicate that the proposed measure can provide accurate predictions of 1) the intelligibility of diotic speech with an accuracy similar to that of the original STOI measure, 2) speech reception thresholds (SRTs) in conditions with a frontal target speaker and a single interferer in the horizontal plane, 3) SRTs in conditions with a frontal target and a single interferer when ideal time frequency segregation (ITFS) is applied to the left and right ears separately, and 4) the advantage of two-microphone beamforming as applied in state-of-the-art hearing aids. A MATLAB implementation of the proposed measure is available online[1].

1 Introduction

The speech intelligibility prediction problem consists of predicting the intelligibility of a particular noisy/processed/distorted speech signal to an average listener. The problem was initially studied with the purpose of improving telephone systems [2, 3]. Since then, it has been applied as a development tool in related fields such as telecommunication [4], architectural acoustics [5, 6] and speech processing [7–12]. Many endeavours in these fields focus on improving speech understanding in particular conditions. This introduces the need for measuring speech intelligibility through listening experiments, which is a time consuming and expensive task. Objective (computational) measures of intelligibility can provide estimates of the results of such experiments faster and at a lower cost, while being easily reproducible.

An early Speech Intelligibility Prediction (SIP) method is the Articulation Index (AI) [3, 13], which can be seen as a common ancestor for most of the methods which have been proposed since then. The AI considers the condition in which a listener is presented with monaural speech contaminated by additive, stationary noise. It is assumed that speech and noise at the ear of the listener are available as separate signals. The AI estimates intelligibility as a weighted sum of normalized Signal to Noise Ratios (SNRs) across a range of third octave bands. It has later been shown that, under certain assumptions,

[1] See `http://kom.aau.dk/project/Intelligibility/`.

this is in fact an estimate of the channel capacity from Shannon's information theory [14]. A refined and standardized version of the AI is known as the Speech Intelligibility Index (SII) [15]. Notably, the AI and the SII are unsuitable for conditions involving fluctuating noise interferers, binaural conditions and conditions where speech and noise are not combined linearly (due to e.g. distorting transmission systems or non-linear speech processing algorithms).

Many SIP methods have been proposed since the introduction of the AI, mainly focussing on extending the domain in which accurate predictions can be made. For example, the Speech Transmission Index (STI) estimates the impact of a transmission channel (e.g. the acoustics of a room or a noisy and distorting transmission system) on intelligibility by measuring the change in modulation depth across the system [16, 17]. It has, however, been shown that the STI does not perform well at predicting the impact of speech enhancement algorithms, on speech intelligibility [9, 18–22]. A more recent modulation-based and physiologically motivated method, the speech-based EPSM (sEPSM), has been shown to perform well at predicting the impact of spectral subtraction [22]. Another notable method is the Extended SII (ESII), which is a variation of the SII that provides more accurate predictions in conditions with fluctuating noise interferers [23, 24]. The Coherence SII (CSII) is yet another variation of the SII which aims to predict the influence on intelligibility of non-linear distortion from clipping [25]. The CSII and several other intelligibility measures are evaluated with speech processed by noise reduction algorithms in [26]. The recent Hearing-Aid Speech Perception Index (HASPI) is closely related to the CSII, but involves a more sophisticated auditory model and aims to predict the intelligibility of processed speech for hearing impaired listers [27]. Recently, the Short-Time Objective Intelligibility (STOI) measure [7] has become very popular for evaluation of noisy and processed speech. The STOI measure has shown to compare favorably to several other SIP methods, with respect to predicting the impact of various single microphone enhancement schemes as well as Ideal Time Frequency Segregation (ITFS) [7]. This observation is confirmed for hearing impaired listeners in [28], which shows that the CSII and the STOI measure perform favorably to other measures at predicting the effect of noise reduction algorithms. The STOI measure has later been shown to compare well with other measures with respect to predicting the impact of a number of detrimental effects and processing schemes relevant to users of hearing aids and cochlear implants [8] and for predicting the intelligibility of noisy speech transmitted by telephone [4]. Finally, we mention the Speech Intelligibility prediction based on Mutual Information (SIMI) measure [12], which is very similar to the STOI measure in structure and performance, but which is based on information theoretical considerations.

The methods discussed up to this point all assume that speech is presented monaurally or diotically to the listener. However, in many real world

scenarios, humans obtain an advantage from listening with two ears. This is partly because one can, to some extent, choose to listen to the ear in which the speech is more intelligible, and partly because the brain can combine information from the two ears [29]. The Equalization Cancellation (EC) stage is an early simple model which predicts Binaural Masking Level Differences (BMLDs) accurately in a range of conditions [30, 31] (i.e. the binaural advantage obtained in tasks such as detecting a tone in noise). Several attempts have been made at developing SIP methods which account for binaural advantage [32–40], i.e. the advantage in intelligibility obtained through the presence of interaural, source-dependent, phase and level differences. Notably, the Binaural Speech Intelligibility Measure (BSIM) [33] uses the EC stage as a preprocessor to the SII to predict the intelligibility of binaural signals. The same paper proposes another binaural method, the short-time BSIM (stBSIM), with properties similar to the ESII (i.e. the ability to handle fluctuating noise interferers) [33]. A number of ways to extend the BSIM, such as to predict the detrimental effect of reverberation, are investigated in [39, 40]. A different approach for combining the SII with an EC stage is proposed in [36]. The method estimates the Speech Reception Threshold (SRT) of the better ear and subtracts an estimate of binaural advantage obtained by an EC based method proposed in [41, 42]. This method has later been expanded further to account for aspects such as to multiple interferers and reverberation [37, 38, 43].

It should be realized that none of the above-mentioned methods are able to predict the simultaneous impact of both non-linear processing and binaural advantage. This is in spite of the fact that both effects are important in the context of modern audio processing devices that present signals dichotically to a user, e.g. hearing aids. In [44] we introduced an early version of the proposed method, that has shown promising results in predicting both the effects of processing and binaural advantage. The method is obtained by extending the STOI measure such as to predict binaural advantage, and is therefore referred to as the Binaural STOI (BSTOI) measure. Taking inspiration from [33], this measure is obtained by using a modified EC stage to combine the left and right ear signals, prior to predicting intelligibility with the STOI measure. Because the EC stage includes internal noise sources, which model inaccuracies in the human auditory system, computationally expensive Monte Carlo simulation is used to obtain an estimate of the expected STOI measure across these noise sources [44]. Results presented in [44] indicate that the BSTOI measure can predict both binaural advantage and the effect of non-linear processing with ITFS. It was not investigated whether the BSTOI measure can account for both effects simultaneously (i.e. they were investigated separately).

In the present study, we introduce a refined version of the BSTOI measure, which we refer to as the Deterministic BSTOI (DBSTOI) measure. In order to avoid Monte Carlo simulation, the DBSTOI measure introduces some

minor changes in the STOI measure which allow us to derive an analytical expression for the expectation of the output measure across the internal noise sources in the EC stage. The DBSTOI measure is therefore much less computationally demanding to evaluate than the BSTOI measure. Furthermore, the DBSTOI measure produces fully deterministic outputs. Except for the mentioned advantages of the DBSTOI measure, no noteworthy performance differences between the DBSTOI and BSTOI measures have been found. Furthermore, we provide a thorough evaluation of the prediction performance of the measure by comparing to the results of four different listening experiments, including one with both non-linear speech enhancement and binaural advantage combined. The ability to handle such conditions allows the measure to predict intelligibility of e.g. users of assistive listening devices in complex real-world scenarios.

The remainder of the paper is organsized as follows: In Sec. 2, the proposed intelligibility measure is described in detail. In Sec. 3, four sets of experimental data are described. In Sec. 4, the procedure used for evaluating the measure is described. In Sec. 5, the results are presented. Sec. 6 concludes upon the proposed measure and its performance.

2 The DBSTOI Measure

In this section we present the proposed intelligibility measure in detail. The measure applies to conditions in which a human subject is listening to a well defined target speaker in the presence of some form of interference. Furthermore, the combination of speech and interferer may have been non-linearly transformed by e.g. a speech enhancement algorithm or a distorting transmission system. Intelligibility is predicted on the basis of four input signals: the left and right *clean* signals, $x_l(t)$ and $x_r(t)$, and the left and right *noisy and processed* signals, $y_l(t)$ and $y_r(t)$. The clean signals are measured at the ear of the listener but in the presence of only the target speaker (and in the absence of both interferer and processing). An example of this is illustrated in Fig. D.1a. It is assumed that the clean signals are fully intelligible. The aim is to predict the intelligibility of the noisy and processed signals. These are given by the processed mixture of target and interferer as measured at the ear of the listener. An example of this is illustrated in Fig. D.1b. The clean and degraded signals are assumed to be time aligned for each ear. E.g. if the degraded signals include a substantial processing delay, the clean signals should be delayed correspondingly to compensate for this difference. It should be stressed that the use of the measure for predicting the impact of hearing aid processed speech, as illustrated in Fig. D.1, is merely an example. The measure is applicable in virtually any condition in which noisy and processed speech is presented binaurally to a listener. The clean and

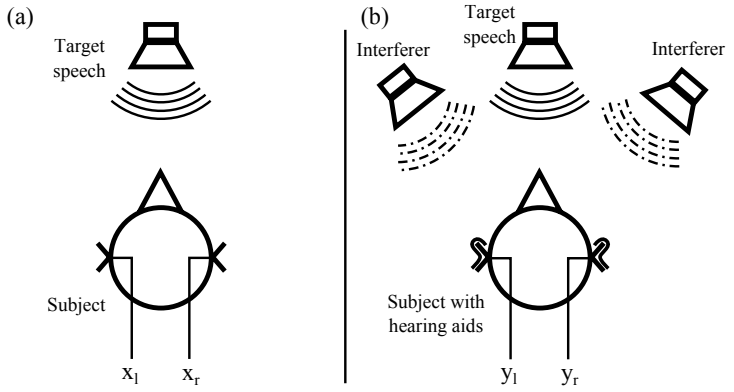


Fig. D.1: The four input signals needed by the proposed measure in the exemplifying application where it is used to predict the intelligibility of speech which has been processed by a hearing aid system. a) The left and right clean signals, $x_l(t)$ and $x_r(t)$, are obtained by measuring the acoustic signal in the ear canal of the left and right ear of the subject when listening only to the unprocessed target speaker. b) The left and right noisy/processed signals, $y_l(t)$ and $y_r(t)$, are measured in the ear canals while the subject is wearing hearing aids and is listening to the combination of target and interferer.

noisy/processed signals may be either recorded or simulated by use of Head Related Transfer Functions (HRTFs). A block diagram of the computational structure of the measure is shown in Fig. D.2. The block diagram is separated in three steps: 1) a Time-Frequency (TF) decomposition based on the Discrete Fourier Transformation (DFT), 2) a modified EC stage which models binaural advantage and 3) a STOI based stage which rates intelligibility on a scale from -1 to $+1$. The three steps are described in detail in sections 2.1, 2.2 and 2.3, respectively.

2.1 Step 1: Time-Frequency Decomposition

The first step of computing the DBSTOI measure is adopted from the STOI measure [7] with no significant changes. The four input signals are first resampled to 10 kHz. Then, regions in which the target speaker is silent are removed with a simple frame-based Voice Activity Detector (VAD). This is done by 1) segmenting the four input signals into 256-sample Hann-windowed segments with an overlap of 50%, 2) finding the frame with the highest energy for each of the two clean signals, respectively, 3) locating all frame indices where the energy of both clean signal frames are more than 40 dB below their respective maximum and 4) resynthesising the four signals, but excluding the frame numbers which where found in 3. This produces four signals which are time aligned, because the same frames are removed in all the signals.

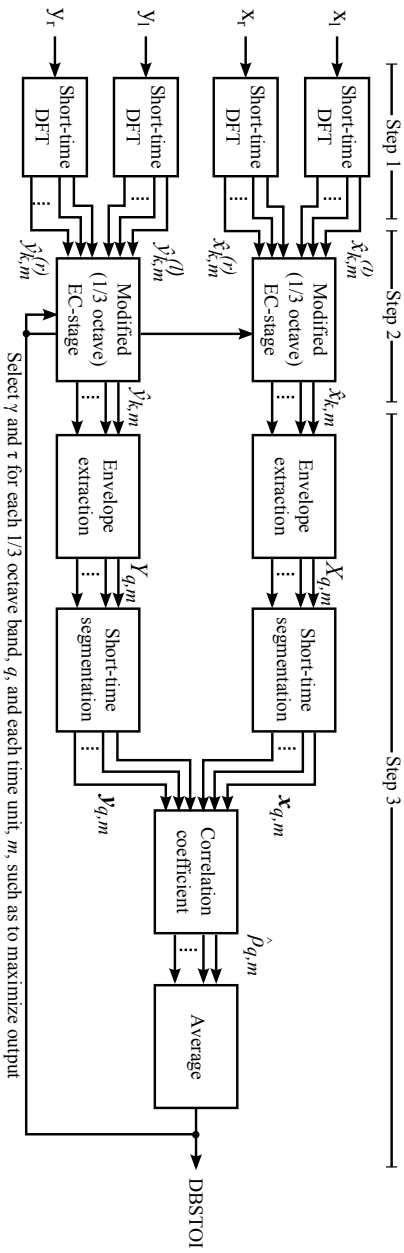


Fig. D.2: A block diagram illustrating the computation of the proposed measure.

A TF decomposition of the signals is then obtained in the same manner as for the STOI measure [7]. This is done by segmenting the signals into 256-sample frames with an overlap of 50%, followed by zero-padding each frame to 512 samples and applying the DFT. We refer to the k'th frequency bin of the m'th frame of the left clean signal as $\hat{x}^{(l)}_{k,m}$. Similarly, the same TF units of the right clean signal and the left and right noisy/processed signals are denoted by $\hat{x}^{(r)}_{k,m}$, $\hat{y}^{(l)}_{k,m}$ and $\hat{y}^{(r)}_{k,m}$, respectively.

2.2 Step 2: Equalization-Cancellation Stage

The second step of computing the measure consists of combining the left and right signals into a single clean signal and a single noisy/processed signal while accounting for any potential binaural advantage. This is done by use of a modified EC stage.

The originally proposed EC stage models binaural advantage under the assumption that the left and right speech and interferer signals are known in separation [30, 31]. The stage introduces relative time shifts and amplitude adjustments between the left and right signals (equalization) and subtracts the two from each other (cancellation) to obtain one signal. This is done separately for the left and right clean signals and the left and right interferer signals such as to obtain a single clean signal and a single interferer signal. The same time shifts and amplitude adjustments are applied for both clean and noisy/processed signals, and these are chosen such as to maximize the SNR of the output. For wideband signals, such as speech, the EC stage is typically applied independently in auditory bands.

The original EC stage cannot be applied in the present case, because the interferer signal is not available in separation. Instead, the processed combination of speech and interferer is available. We propose the following changes in order to adapt the EC stage to work with the available signals:

1. The left and right clean signals and the left and right noisy/processed signals are combined using the same procedure as that of the original EC stage.

2. The time shifts and amplitude adjustment factors are determined such as to maximize the STOI measure of the output, rather than the SNR.

This essentially correspond to assuming that the human brain combines the signals from the two ears such as to maximize intelligibility rather than SNR. The combination of the left and right signals by the modified EC stage, is carried out in the frequency domain as follows:

$$\hat{x}_{k,m} = \lambda_{k,m}\, \hat{x}^{(l)}_{k,m} - \lambda^{-1}_{k,m} \hat{x}^{(r)}_{k,m}, \tag{D.1}$$

$$\hat{y}_{k,m} = \lambda_{k,m}\, \hat{y}^{(l)}_{k,m} - \lambda^{-1}_{k,m} \hat{y}^{(r)}_{k,m}, \tag{D.2}$$

where the time and frequency dependent complex-valued factor $\lambda_{k,m}$ represents the time shift and the amplitude adjustment. Specifically, this factor is given by:

$$\lambda_{k,m} = 10^{(\gamma_{k,m}+\Delta\gamma_{k,m})/40} e^{j\omega(\tau_{k,m}+\Delta\tau_{k,m})/2}, \tag{D.3}$$

where $\gamma_{k,m}$ is the relative amplitude adjustment (in dB), $\tau_{k,m}$ is the relative time shift (in seconds), and $\Delta\gamma_{k,m}$ and $\Delta\tau_{k,m}$ are uncorrelated random variables which serve to model the suboptimal performance of the human auditory system [30, 45]. These are normally distributed with zero mean and variance (adapted from [45] in the same manner as is done in [32, 33]) given by[2]:

$$\sigma_{\Delta\gamma}(\gamma_{k,m}) = \sqrt{2} \cdot 1.5 \text{ dB} \cdot \left(1 + \left(\frac{|\gamma_{k,m}|}{13 \text{ dB}}\right)^{1.6}\right) \quad [\text{dB}], \tag{D.4}$$

$$\sigma_{\Delta\tau}(\tau_{k,m}) = \sqrt{2} \cdot 65 \ \mu\text{s} \cdot \left(1 + \frac{|\tau_{k,m}|}{1.6 \text{ ms}}\right) \quad [\text{s}]. \tag{D.5}$$

The values of $\gamma_{k,m}$ and $\tau_{k,m}$ are determined independently for each time unit and third octave band such as to maximize the STOI measure of the combined signals (i.e. $\gamma_{k,m}$ and $\tau_{k,m}$ have the same value for all k belonging to one third octave band). The details of this are covered in Sec. 2.4. Henceforth, for notational convenience, we discard time and frequency indices such as to denote $\lambda_{k,m}$, $\gamma_{k,m}$ and $\tau_{k,m}$ as λ, γ and τ, respectively. The same is done for the noise sources $\Delta\gamma$ and $\Delta\tau$.

2.3 Step 3: Intelligibility Prediction

At this point, the left and right ear signals have been combined into one clean signal, $\hat{x}_{k,m}$, and one noisy/processed signal, $\hat{y}_{k,m}$, cf. Fig. D.2. This allows us to estimate intelligibility using the STOI measure. However, the signals $\hat{x}_{k,m}$ and $\hat{y}_{k,m}$ are stochastic due to the noise sources $\Delta\gamma$ and $\Delta\tau$ in the EC stage, and the resulting STOI measure is therefore also a stochastic variable. For the BSTOI measure this problem was solved by averaging the output across many realizations of $\Delta\gamma$ and $\Delta\tau$ [44]. This solution is computationally expensive and does not lead to entirely deterministic results. The DBSTOI measure instead applies a slight variation of the originally proposed STOI measure[3],

[2] In [45], noise sources are added independently to the left and right ear signals. Here, one noise source is applied symmetrically. This leads to a multiplicative factor of $\sqrt{2}$ in (D.4) and (D.5) compared to [45].

[3] For mathematical tractability, we use "power envelopes" (envelopes squared) rather than magnitude envelopes as originally proposed for the STOI measure [7]. This is also done in [46] and appears to have no significant effect on predictions [46, 47]. Furthermore, we discard the clipping mechanism used in the original STOI measure. The same variation is applied in [46] and the changes do not appear to significantly impair the prediction performance of the measure.

which allows us to derive a closed form expression of the expectation of the final measure across $\Delta\gamma$ and $\Delta\tau$. The remainder of this section describes these matters in detail.

The clean signal "power envelopes" (envelopes squared) are first determined in $Q = 15$ third octave bands with center frequencies starting from 150 Hz. These bands are obtained by grouping DFT coefficients, exactly as in the original STOI measure [7]. The border between two adjacent bands are given by the geometric mean of their respective center frequencies. The upper and lower frequency bin indices of the q'th band are denoted, respectively, by $k_1(q)$ and $k_2(q)$. The resulting expression for the clean signal power envelope is given by:

$$\begin{aligned} X_{q,m} &= \sum_{k=k_1(q)}^{k_2(q)} \left|\hat{x}_{k,m}\right|^2 = \sum_{k=k_1(q)}^{k_2(q)} \left|\lambda\, \hat{x}_{k,m}^{(l)} - \lambda^{-1}\hat{x}_{k,m}^{(r)}\right|^2 \\ &= 10^{\frac{\gamma+\Delta\gamma}{20}} \sum_{k=k_1(q)}^{k_2(q)} \left|\hat{x}_{k,m}^{(l)}\right|^2 + 10^{-\frac{\gamma+\Delta\gamma}{20}} \sum_{k=k_1(q)}^{k_2(q)} \left|\hat{x}_{k,m}^{(r)}\right|^2 \\ &\quad - 2\mathrm{Re}\left[\sum_{k=k_1(q)}^{k_2(q)} \hat{x}_{k,m}^{(l)*}\hat{x}_{k,m}^{(r)} e^{-j\omega_k(\tau+\Delta\tau)}\right] \\ &\approx 10^{\frac{\gamma+\Delta\gamma}{20}} X_{q,m}^{(l)} + 10^{-\frac{\gamma+\Delta\gamma}{20}} X_{q,m}^{(r)} \\ &\quad - 2\,\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)} X_{q,m}^{(c)}\right], \end{aligned} \tag{D.6}$$

where:

$$\begin{aligned} X_{q,m}^{(l)} &= \sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m}^{(l)}|^2, \\ X_{q,m}^{(r)} &= \sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m}^{(r)}|^2, \\ X_{q,m}^{(c)} &= \sum_{k=k_1(q)}^{k_2(q)} \hat{x}_{k,m}^{(l)*}\hat{x}_{k,m}^{(r)}, \end{aligned} \tag{D.7}$$

and where ω_k is the angular frequency of the k'th frequency bin and ω_q is the center angular frequency of the q'th third octave band. The last step in (D.6) assumes that the signal energy is located at the center of each third octave band. The same procedure is applied for the noisy/processed signal to obtain $Y_{q,m}$ as well as $Y_{q,m}^{(l)}$, $Y_{q,m}^{(r)}$ and $Y_{q,m}^{(c)}$.

The obtained power envelope samples are then arranged temporally in zero-mean vectors of $N = 30$ samples, in the same manner as is done in

the STOI measure [7]:

$$\mathbf{x}_{q,m} = [X_{q,m-N+1}, \ldots, X_{q,m}]^\mathsf{T} - \mathbf{1} \sum_{m'=m-N+1}^{m} \frac{X_{q,m'}}{N}, \tag{D.8}$$

where $\mathbf{1}$ is a column vector of all ones. Similar vectors are defined from the other power envelope signals: $\mathbf{x}^{(l)}_{q,m}$, $\mathbf{x}^{(r)}_{q,m}$, $\mathbf{x}^{(c)}_{q,m}$, $\mathbf{y}^{(l)}_{q,m}$, $\mathbf{y}^{(r)}_{q,m}$ and $\mathbf{y}^{(c)}_{q,m}$. From (D.6) we then have:

$$\begin{aligned}\mathbf{x}_{q,m} \approx\ & 10^{\frac{\gamma+\Delta\gamma}{20}} \mathbf{x}^{(l)}_{q,m} + 10^{-\frac{\gamma+\Delta\gamma}{20}} \mathbf{x}^{(r)}_{q,m} \\ & - 2\,\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)} \mathbf{x}^{(c)}_{q,m}\right].\end{aligned} \tag{D.9}$$

A similar expression holds for $\mathbf{y}_{q,m}$.

In order to compute the expectation across the final measure, we assume that the input signals are wide sense stationary stochastic processes across the duration of one segment (i.e. 386 ms). It follows that the third octave band envelope samples, $X_{q,m}$ and $Y_{q,m}$ are also samples of a stochastic process, due to the stochastic nature of the input signals, but also due to the random variables, $\Delta\gamma$ and $\Delta\tau$, introduced in the EC stage. A basic assumption of the original STOI measure is that speech intelligibility is related to the average sample correlation between the vectors $\mathbf{x}_{q,m}$ and $\mathbf{y}_{q,m}$ [7]. This, however, may be interpreted simply as an estimate of the correlation between the processes $X_{q,m}$ and $Y_{q,m}$:

$$\rho_{q,m} = \frac{E\left[(X_{q,m} - E\left[X_{q,m}\right])(Y_{q,m} - E\left[Y_{q,m}\right])\right]}{\sqrt{E\left[(X_{q,m} - E\left[X_{q,m}\right])^2\right] E\left[(Y_{q,m} - E\left[Y_{q,m}\right])^2\right]}}, \tag{D.10}$$

where the expectation is taken across both input signals, $\Delta\gamma$ and $\Delta\tau$. An estimate of this expectation across $N = 30$ envelope samples is given by:

$$\bar{\rho}_{q,m} = \frac{E_\Delta\left[\mathbf{x}^\mathsf{T}_{q,m}\mathbf{y}_{q,m}\right]}{\sqrt{E_\Delta\left[||\mathbf{x}_{q,m}||^2\right] E_\Delta\left[||\mathbf{y}_{q,m}||^2\right]}}, \tag{D.11}$$

where $E_\Delta[\cdot]$ denotes the expectation across $\Delta\gamma$ and $\Delta\tau$. A closed form ap-

proximation of this expectation is derived in Appendix A, and is given by:

$$\begin{aligned}
&E_{\Delta}\left[\mathbf{x}_{q,m}^{\mathsf{T}}\mathbf{y}_{q,m}\right] \approx \\
&\quad (e^{2\beta}\mathbf{x}_{q,m}^{(l)\mathsf{T}}\mathbf{y}_{q,m}^{(l)} + e^{-2\beta}\mathbf{x}_{q,m}^{(r)\mathsf{T}}\mathbf{y}_{q,m}^{(r)})e^{2\sigma_{\Delta\beta}^2} \\
&+ \mathbf{x}_{q,m}^{(r)\mathsf{T}}\mathbf{y}_{q,m}^{(l)} + \mathbf{x}_{q,m}^{(l)\mathsf{T}}\mathbf{y}_{q,m}^{(r)} - 2e^{\sigma_{\Delta\beta}^2/2}e^{-\omega^2\sigma_{\Delta\tau}^2/2}\times \\
&\quad \left\{\left(e^{\beta}\mathbf{x}_{q,m}^{(l)\mathsf{T}} + e^{-\beta}\mathbf{x}_{q,m}^{(r)\mathsf{T}}\right) Re\left[\mathbf{y}_{q,m}^{(c)}e^{-j\omega\tau}\right]\right. \\
&+ \left. Re\left[e^{-j\omega\tau}\mathbf{x}_{q,m}^{(c)}\right]^{\mathsf{T}}\left(e^{\beta}\mathbf{y}_{q,m}^{(l)} + e^{-\beta}\mathbf{y}_{q,m}^{(r)}\right)\right\} \\
&+ 2\left(Re\left[\mathbf{x}_{q,m}^{(c)H}\mathbf{y}_{q,m}^{(c)}\right] + e^{-2\omega^2\sigma_{\Delta\tau}^2}Re\left[\mathbf{x}_{q,m}^{(c)\mathsf{T}}\mathbf{y}_{q,m}^{(c)}e^{-j2\omega\tau}\right]\right),
\end{aligned} \tag{D.12}$$

where:

$$\begin{aligned}
\beta &= \frac{\ln(10)}{20}\gamma, \\
\sigma_{\Delta\beta}^2 &= \left(\frac{\ln(10)}{20}\right)^2\sigma_{\Delta\gamma}^2.
\end{aligned} \tag{D.13}$$

The approximation in (D.12) stems from the approximation introduced in (D.6).

A similar expression can be used to compute $E_{\Delta}\left[||\mathbf{x}_{q,m}||^2\right] = E_{\Delta}\left[\mathbf{x}_{q,m}^{\mathsf{T}}\mathbf{x}_{q,m}\right]$. This is obtained simply by replacing all occurrences of $\mathbf{y}$ in (D.12) with $\mathbf{x}$. In a similar manner, an expression for $E_{\Delta}\left[||\mathbf{y}_{q,m}||^2\right]$ can be obtained. This makes it possible to evaluate (D.11) in closed form.

In the same manner as in the STOI measure, we define the final measure to be the average of these correlation estimates:

$$\text{DBSTOI} = \frac{1}{QM}\sum_{m=1}^{M}\sum_{q=1}^{Q}\bar{\rho}_{q,m}. \tag{D.14}$$

It should be noted that $\bar{\rho}_{q,m}$ is dependent on the parameters of the EC stage, γ and τ.

2.4 Determination of γ and τ

As stated, the parameters γ and τ are determined such as to maximize predicted intelligibility, i.e. (D.14). These parameters are determined independently for each estimated correlation coefficient, $\bar{\rho}_{q,m}$, i.e. for each time unit m and third octave band q. The values of $\gamma = \gamma_{k,m}$ and $\tau = \tau_{k,m}$ are therefore held constant for all frequency bins, k, within one third octave band, q, and for all N envelope samples, m, within one set of envelope vectors, $\{\mathbf{x}_{q,m},\ \mathbf{y}_{q,m}\}$.

The values of γ and τ are found separately for each estimated correlation coefficient, by maximizing the correlation:

$$\bar{\rho}_{q,m} = \max_{\gamma,\tau}\ \bar{\rho}_{q,m}(\gamma,\tau). \tag{D.15}$$

Table D.1: Time spent producing a scoring of 100 seconds of white noise on a Lenovo W530 with an Intel Core i7-3820QM, 2.7 GHz. The authors' own MATLAB implementation of the DBSTOI measure was used, while a STOI measure implementation was provided by the authors of [7].

Algorithm	STOI	DBSTOI
Time	5.3 s	62.2 s

It has not been possible to find a simple analytical procedure for solving this optimization problem. Instead, an approximately optimal solution is found by evaluating (D.15) for a range of combinations of γ and τ. In practice, we search all combinations of an evenly spaced range of 100 τ-values from -1 ms to $+1$ ms and an evenly spaced range of 40 γ values from -20 dB to $+20$ dB. On top of the mentioned (γ,τ)-combinations, the correlation is also estimated for each ear individually (a "better-ear option"), corresponding to $\gamma = \pm\infty$. These estimates are as well included in the search of the maximum in (D.15). Preliminary experiments have indicated that the quality of the output is not highly sensitive to the searched range of (γ,τ)-combinations. For applications with scarce computational resources, the cost of computing the measure can be lowered by more coarsely searching these variables. The computational cost of the DBSTOI measure is compared to that of the STOI measure in Table D.1. It should, however, be noted that the cost of both methods depends on the choice of parameters and can most likely be decreased significantly by implementing them a low-level language.

A noteworthy special case of the DBSTOI measure arises for diotic stimuli, where $\mathbf{x}_{q,m}^{(l)} = \mathbf{x}_{q,m}^{(r)} = \mathbf{x}_{q,m}^{(c)}$ and $\mathbf{y}_{q,m}^{(l)} = \mathbf{y}_{q,m}^{(r)} = \mathbf{y}_{q,m}^{(c)}$. This implies that $\mathbf{x}_{q,m}^{(c)}$ and $\mathbf{y}_{q,m}^{(c)}$ are real-valued. Therefore, (D.12) simplifies to:

$$\begin{aligned} E_\Delta\left[\mathbf{x}_{q,m}^\intercal \mathbf{y}_{q,m}\right] \approx & \left(\left(e^{2\beta}+e^{-2\beta}\right)e^{\sigma_{\Delta\beta}^2}+2\right. \\ & -4 2 e^{\sigma_{\Delta\beta}^2/2} e^{-\omega^2\sigma_{\Delta\tau}^2/2}\left(e^{\beta}+e^{-\beta}\right) Re\left[e^{-j\omega\tau}\right]+2 \\ & \left.+e^{-2\omega^2\sigma_{\Delta\tau}^2} Re\left[e^{-j2\omega\tau}\right]\right)\mathbf{x}_{q,m}^{(l)\intercal}\mathbf{y}_{q,m}^{(l)}. \end{aligned} \tag{D.16}$$

Inserting this in (D.11), it can be verified that the entire τ- and β-dependent factor cancels, because the same factor appears in the denominator. This implies that the DBSTOI measure simplifies to the monaural (modified) STOI measure for diotic signals.

3 Experimental Data

We evaluate the performance of the proposed measure by comparing it to the results of four listening experiments. In this section we describe these experiments.

The first two experiments make it possible to investigate the performance of the proposed measure in comparison with other measures of intelligibility. The third and fourth experiments make it possible to investigate the performance of the DBSTOI measure in conditions with both binaural advantage and processing. The conditions of experiments 2–4 are summarized in Table D.2 (Experiment 1 is excluded due to the large number of conditions and the fact that it is thoroughly documented in [48]).

3.1 Experiment 1: Diotic Presentation and Ideal Time Frequency Segregation

This data set was collected as part of a study on ITFS [48], but has kindly been made available for evaluation of the present work. Subjects were presented with noisy and processed sentences from the Dantale II corpus [49]. After each sentence, the subjects were requested to repeat as many words as possible. The experimenter marked the correctly repeated words. Sentences were presented diotically via headphones together with one of four different interferers: Speech Shaped Noise (SSN), café noise, bottling factory noise and car interior noise. Sentences mixed with each noise type were ITFS processed with Ideal Binary Masks (IBMs) at 8 different threshold values. Furthermore, sentences mixed with each noise type, excluding SSN, were ITFS processed with Target Binary Masks (TBMs) at 8 different threshold values. Each combination of noise and processing was presented at 3 SNRs and with two sentences for each. The experiment was carried out with 15 normal hearing Danish speaking subjects. This resulted in the collection of results for 15 subjects $\times$ 7 noise/mask combinations $\times$ 8 thresholds $\times$ 3 SNRs $\times$ 2 repetitions $=$ 5040 sentences. See [48] for a detailed description of the experimental procedure.

The original STOI measure has been shown to correlate well with the results of this experiment [7]. We include it in this study in order to investigate the impact of the differences between the STOI and DBSTOI measures for diotic stimuli.

3.2 Experiment 2: A Single Source of SSN in the Horizontal Plane

Speech intelligibility was measured in the condition of a frontal speaker masked by a single SSN interferer in the horizontal plane [44]. An anechoic environment was simulated binaurally by use of the CIPIC HRTFs [50] and the result was presented via Sennheiser HDA200 headphones at a comfortable level. Sentences from the Danish Dantale II material were used as target signals while SSN was generated by filtering Gaussian noise to have the same long time spectrum as these sentences. Speech intelligibility was

Table D.2: Summary of experiments 2–4. For details see the text.

Cond.	Interferer type	Interferer location	Proc.
2.1	SSN	$-160°$	–
2.2	SSN	$-115°$	–
2.3	SSN	$-80°$	–
2.4	SSN	$-40°$	–
2.5	SSN	$20°$	–
2.6	SSN	$0°$	–
2.7	SSN	$40°$	–
2.8	SSN	$80°$	–
2.9	SSN	$140°$	–
2.10	SSN	$180°$	–
3.1	SSN	$-115°$	ITFS
3.2	SSN	$0°$	ITFS
3.3	SSN	$20°$	ITFS
3.4	Bottling factory	$-115°$	–
3.5	Bottling factory	$0°$	–
3.6	Bottling factory	$20°$	–
3.7	Bottling factory	$-115°$	ITFS
3.8	Bottling factory	$0°$	ITFS
3.9	Bottling factory	$20°$	ITFS
4.1	SSN	isotropic	–
4.2	SSN	$\{-115°,180°,115°\}$	–
4.3	ISTS	$\{-115°,180°,115°\}$	–
4.4	SSN	$\{30°,180°\}$	–
4.5	ISTS	$\{30°,180°\}$	–
4.6	SSN	isotropic	Beamforming
4.7	SSN	$\{-115°,180°,115°\}$	Beamforming
4.8	ISTS	$\{-115°,180°,115°\}$	Beamforming
4.9	SSN	$\{30°,180°\}$	Beamforming
4.10	ISTS	$\{30°,180°\}$	Beamforming

measured for 10 interferer angles, each for 6 SNRs. The SNRs were equally spaced by 3 dB, centred around a rough estimate of the SRT for each condition. Sentences were presented one at a time and the subject was requested to repeat as many words of each sentence as possible. The experimenter marked the correctly repeated words. Three sentences were presented for each combination of interferer position and SNR. The experiment was carried out for 10 normal hearing Danish speaking subjects. In total, results were collected from the presentation of 10 subjects $\times$ 10 interferer positions $\times$ 6 SNRs $\times$ 3 repetitions $=$ 1800 sentences. The conditions of Experiment 2 are summarized in Table D.2.

The conditions of Experiment 2 contain no processing and are therefore applicable to a range of existing binaural SIP methods. We include it in this study to allow for a comparison of the DBSTOI measure with existing measures.

3.3 Experiment 3: A Single Interferer in the Horizontal Plane and Ideal Time Frequency Segregation

This experiment measured intelligibility in 9 conditions with a frontal speaker masked by a single interferer. Conditions 1–3 included an SSN masker at some position in the horizontal plane (as in Experiment 2) but with ITFS applied independently to the signals of each ear. This was done in as manner similar to that described in [51]: 1) a short-time DFT was applied to the speech and interferer signals in separation, prior to mixing, 2) DFT coefficients for which the SNR of the mixed signal was below 0 dB (i.e. where the magnitude of the interferer coefficient was larger than that of the target coefficient) were attenuated by 10 dB, 3) the signals were reconstructed. The finite attenuation of 10 dB was chosen to restrict the improvement in intelligibility (as ITFS can, otherwise, make speech fully intelligible regardless of SNR [48]). The interferer positions were chosen as a representative subset of those used in Experiment 2. The same interferer positions were used for conditions 4–6 and 7–9. In conditions 4–6, a "bottling factory noise" was used as interferer in place of SSN. This is a fluctuating noise type with more energy at higher frequencies than speech [52]. Conditions 7–9 were the same as 4–6 but with ITFS. The Dantale II corpus was also used in this experiment. The environment was simulated using the CIPIC HRTFs [50] and the signals presented via Sennheiser HDA200 headphones. The subjects were presented with one sentence at a time. After each sentence the subjects were requested to select the words they heard on a screen. For each of the five words in each sentence the subjects were offered a choice between 10 possible words and the option to pass (if the word had not been heard at all). In [53, 54], this procedure is shown to yield results almost identical to the verbal procedure used for collecting the results

of Experiment 2. The experiment was run with 14 normal hearing Danish speaking subjects. In total, results were collected from the presentation of 14 subjects $\times$ 9 conditions $\times$ 6 SNRs $\times$ 3 repetitions $=$ 2268 sentences. The conditions of Experiment 3 are summarized in Table D.2.

While experiments 1 and 2 investigate conditions with *either* processing or binaural advantage, Experiment 3 includes conditions with *both* processing and binaural advantage. Therefore, none of the mentioned existing SIP methods can be applied.

3.4 Experiment 4: Multiple Interferers and Beamforming

This experiment measured speech intelligibility in 10 somewhat more complex conditions relevant to the evaluation of hearing aids [44]. Conditions were again simulated binaurally by use of HRTFs and presented via Sennheiser HD 280 Pro headphones at a comfortable level. The Dantale II speech material was used as target speech material and the target speaker was placed in front of the subject in all conditions. Responses were given in the same way as in Experiment 2. In each condition, the subject was presented with speech at 6 different SNRs. In condition 1, the target was masked by cylindrically isotropic SSN. In condition 2 the target was masked by three sources of SSN positioned in the horizontal plane at azimuths of 110°, 180° and -110°. Condition 3 was the same as condition 2, but used randomly selected segments of the International Speech Test Signal (ISTS) [55] instead of SSN as interferer. In condition 4 the target was masked by two sources of SSN positioned in the horizontal plane at azimuths of 30° and 180°. Condition 5 was the same as condition 4, but again used segments of the ISTS instead of SSN as interferer. Conditions 6–10 were the same as conditions 1–5 but included 2-microphone beamforming as used in hearing aids. This was accomplished by using HRTFs measured from far field and to the two microphones of a behind-the-ear hearing aid, and combining the obtained signals with a time-invariant linear MVDR beamformer. The experiment was carried out with 10 normal hearing Danish speaking subjects. In total, results were collected from the presentation of 10 subjects $\times$ 10 conditions $\times$ 6 SNRs $\times$ 3 repetitions $=$ 1800 sentences. The conditions of Experiment 4 are summarized in Table D.2.

Experiment 4 is included to provide insights into the performance of the DBSTOI measure in acoustically varied scenes. Furthermore, beamforming is an increasingly important type of processing in e.g. hearing devices.

4 Evaluation Procedure

Sec. 3 presents a substantial quantity of data. In each of the conditions, considered in the experiments, we can rate the intelligibility on an arbitrary scale

(i.e. one that has an unknown relationship with speech intelligibility), using the proposed DBSTOI measure. In this section we present a range of tools which are used to compare the experimental results to these objective ratings of intelligibility, and thereby to quantify the performance of the DBSTOI measure.

4.1 Representation of Experimental Data

We represent the results of the described listening experiments either in terms of the average fraction of correctly repeated words or in terms of SRTs. We define the SRT as the SNR at which a subject is able to correctly repeat 50% of words. We determine this point from the measured data by maximum-likelihood-fitting a logistic function [56]:

$$p(\mathrm{SNR}) = \frac{1}{1 + e^{4 \cdot s_0 \cdot (\mathrm{SRT} - \mathrm{SNR})}}, \tag{D.17}$$

with respect to the parameters SRT and s_0, where s_0 is the slope of the function at SNR = SRT.

4.2 Predicting the Fraction of Correct Words

The DBSTOI measure provides an output on an arbitrary scale. We assume that a monotonic relationship exists between the DBSTOI measure and actual intelligibility (i.e. the fraction of words repeated correctly). In the proposal of the STOI measure, a logistic function was used to model this relationship [7]. The same procedure is followed here:

$$f(d) = \frac{100\%}{1 + e^{ad+b}}, \tag{D.18}$$

where d is the DBSTOI measure, $f(d)$ is the estimated fraction of correctly repeated words and a and b are free parameters which we fit by maximum likelihood, such as to provide the best possible predictions.

4.3 Prediction of SRTs

By calibrating the proposed measure to a reference condition with known SRT, we may directly predict SRTs for other conditions. First, the proposed measure is evaluated for the reference condition at SRT, in order to output a reference value. Assuming that the measure correlates well with intelligibility, this reference value can be assumed to correspond to the SRT in other conditions as well. We may therefore predict the SRT for another condition by evaluating the proposed measure for a sequence of different input SNRs which are chosen adaptively such that the output approaches the reference

value (e.g. using bisection). The SNR at which this procedure converges is taken to be an estimate of the SRT.

4.4 Measures of Prediction Accuracy

Whenever comparing listening test results and corresponding predictions, we rely on the following three performance statistics. Let x_i be the experimentally measured intelligibility (either fraction of correctly repeated words or the SRT) and y_i be corresponding predicted intelligibility, for conditions $i = 1, \ldots, I$. The performance statistics are then given by:

1. Sample standard deviation:

$$\sigma = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I}(y_i - x_i)^2}. \tag{D.19}$$

2. Pearson correlation:

$$\rho = \frac{\sum_{i=1}^{I}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{I}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{I}(y_i - \mu_y)^2}}, \tag{D.20}$$

   where $\mu_x = \frac{1}{I}\sum_{i=1}^{I} x_i$ and $\mu_y = \frac{1}{I}\sum_{i=1}^{I} y_i$.

3. Kendall rank correlation [57][4]:

$$\phi = \frac{c_\mathrm{c} - c_\mathrm{d}}{\frac{1}{2}I(I-1)}, \tag{D.21}$$

   where c_c is the number of concordant pairs, i.e. the number of unique tuples, (i, j), such that $(x_i > x_j) \wedge (y_i > y_j) \vee (x_i < x_j) \wedge (y_i < y_j)$, and c_d is the number of discordant pairs, i.e. the number of unique tuples, (i, j), such that $(x_i > x_j) \wedge (y_i < y_j) \vee (x_i < x_j) \wedge (y_i > y_j)$.

5 Results

In this section we apply the proposed measure to yield predictions of the results of the listening experiments described in Sec. 3.

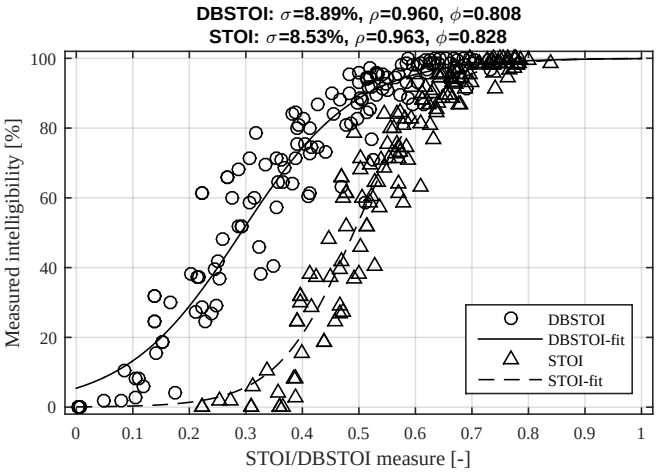


Fig. D.3: Measured intelligibility for the conditions/SNRs of Experiment 1 compared with the DBSTOI measure and the STOI measure. A logistic function was maximum likelihood fitted for each method. The statistics σ and ρ were computed by comparing the shown data points with the predictions made by the fitted logistic curves.

5.1 Diotic Presentation and Ideal Time Frequency Segregation

We first consider the data of Experiment 1 (Sec. 3.1), which investigated the intelligibility of diotic noisy speech processed with ITFS. Due to the diotic nature of the signals, we can obtain predictions by use of the original STOI measure. The STOI measure was designed with the main focus of predicting the impact of TF weighting such as ITFS, and has been shown to perform very well at doing so [7]. We may also obtain predictions with the DBSTOI measure, by simply using the same signals as inputs to the left and right channels of the measure (corresponding to presenting the same signals on the left and right ears of a subject). This allows us to investigate whether the desirable performance of the STOI measure is retained in the DBSTOI measure in spite of the introduced modifications.

Predictions with both the STOI and DBSTOI measures were based on sequences of 30 Dantale sentences. Measured intelligibility was obtained for each condition/SNR by averaging the fraction of correct words across subjects and repetitions. The results are shown in Fig. D.3. It is evident that there is a strong relationship between measured intelligibility and both STOI and DBSTOI measures. The three statistics, shown at the top of Fig. D.3, indicate that the STOI measure performs marginally better than the proposed

[4]Conventionally, τ is used for the Kendall rank correlation. We do not follow this convention because τ is used for a different purpose throughout the paper.

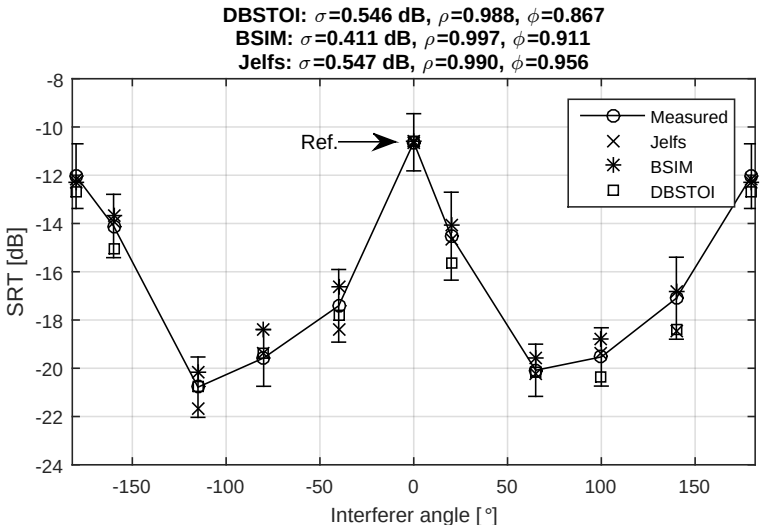


Fig. D.4: SRTs estimated from the results of Experiment 2 along with the predictions of three different measures of speech intelligibility. The measures are all calibrated to the S_0N_0-condition (where both speech and interference comes from the front). The error bars show the standard deviation of SRTs among subjects for the measured results. The reference condition is marked by an arrow.

measure. In Sec. 2.4, it was shown that the effect of the EC stage cancels for diotic stimuli. Therefore, the differences between the STOI measure and the DBSTOI measure, in these conditions, stem only from the modifications introduced in the STOI measure, and not from the extension with an EC stage[5]. The similar performance of the two methods indicate that the modifications applied to the STOI measure to obtain the proposed measure do not strongly impair the performance in a task which is central to the original STOI measure.

5.2 A Single Source of SSN in the Horizontal Plane

We now consider the results of Experiment 2 (Sec. 3.2), which involved frontal speech masked by a single source of SSN in the horizontal plane. The conditions of this experiment allow subjects to obtain a binaural advantage but include no processing. The STOI measure is unsuitable for predicting these results as it relies on monaural/diotic signals. Instead we compare the predictions of the DBSTOI measure to predictions of two existing methods which do consider binaural advantage (but which do not allow for non-linearly pro-

[5]The slight decrease in performance may stem from either the use of "power envelopes" rather than conventional envelopes, or from the fact that the DBSTOI measure does not include a clipping mechanism such as the one used in the original STOI [7].

cessed signals).

Firstly, we compare to the BSIM [33], which is a binaural measure obtained by combining the EC stage with the SII. The BSIM requires knowledge of binaural speech and interference in separation and outputs a number between 0 and 1. It can be used to predict SRTs following the procedure described in Sec. 4.3. This was done by calibrating to the S_0N_0-condition (the condition in which speech and interference sources are co-located in front of the listener). When carrying out predictions with the BSIM, SSN was used for both target and interferer signals (as the method is also evaluated in this manner in [33]). The BSIM was implemented following the description given in [33].

Secondly, we compare to a method described by Jelfs et. al. in [37]. This method uses an SII-like scheme to predict the SRT for the better ear, and a correlation-based model for predicting the additional binaural advantage. The method outputs SRTs but with a significant offset [37], and was therefore calibrated by shifting all outputs by an additive constant, chosen such as to yield correct predictions in the S_0N_0-condition. For this method, SSN was also used as both target and masker (as the method is also evaluated in this manner in [37]). An implementation of this method was kindly provided by the authors of [37].

The DBSTOI measure was used to carry out SRT predictions as described in Sec. 4.3 after being calibrated to the S_0N_0-condition. The predictions were based on a clean signal composed of 30 concatenated Dantale II sentences and an SSN interferer of the same length, both convolved with appropriate HRTFs. Signals of the same length were used for the methods of comparison.

Fig. D.4 shows the results of measurements along with the predictions of the three methods. It is evident that all methods produce very accurate predictions in all conditions, especially considering the standard deviations on the measurements, and the fact that measurements of binaural advantage can vary by several dB from one study to another [29]. The statistics on top of Fig. D.4 indicate that the DBSTOI measure produces predictions which are slightly less accurate than those of the BSIM. This conclusion, however, should be viewed in the light of the facts that 1) the DBSTOI measure does not have access to the interferer in separation (i.e. it does not assume that the speech signal is merely degraded by an additive interferer), while the two existing measures require access to speech and interference in separation, and 2) the DBSTOI measure uses actual speech signals as input while the two existing measures use SSN as both speech and interferer signals.

The DBSTOI measure predictions in isolation, indicate that the measure can indeed predict binaural advantage. This suggests that the modifications introduced in the EC stage, to make it handle non-linearly processed signals, have not severely degraded its performance.

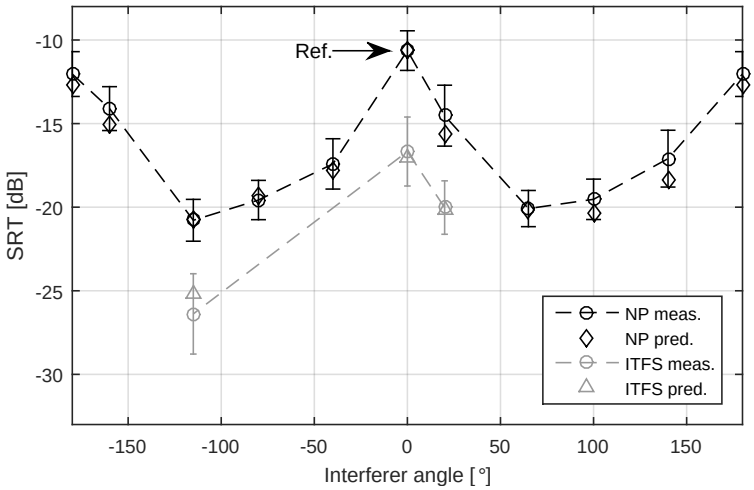


Fig. D.5: Comparison between measured and predicted SRTs for all 10 conditions of Experiment 2 (denoted NP) and conditions 1–3 from Experiment 3 (denoted ITFS). The conditions are shown together because they all use SSN for masking. SRTs were estimated from the measured results for each subject individually for each condition. The results, averaged across subjects, are shown as dotted lines with standard deviations. The SRT of one reference condition was used to make SRT predictions for the other conditions. The reference condition is marked by an arrow.

Table D.3: Predicted and measured binaural advantage.

	Meas. bin. adv.	Pred. bin.adv.
SSN-NP	10.2 dB	10.2 dB
SSN-ITFS	9.7 dB	8.2 dB
BFN-NP	8.8 dB	10.2 dB
BFN-ITFS	7.1 dB	6.1 dB

5.3 A Single Interferer in the Horizontal Plane and Ideal Time Frequency Segregation

In this section we consider Experiment 3 (Sec. 3.3), which is similar to Experiment 2, but involves conditions with ITFS and masking by either SSN or bottling factory noise (see Table D.2). We remark that the STOI measure is not applicable to predicting the intelligibility of the conditions in Experiment 3, as binaural advantage is involved. At the same time, the BSIM or the method by Jelfs et. al. are not applicable, as non-linear processing in the form of ITFS is involved. We therefore present results from the DBSTOI measure alone. This serves as a study of the prediction performance of the DBSTOI measure in conditions with *both* binaural advantage *and* processing. Predictions of SRTs were carried out as described in Sec. 4.3, with calibration to the same condition as in Sec. 5.2. The predictions were based on 30 concatenated

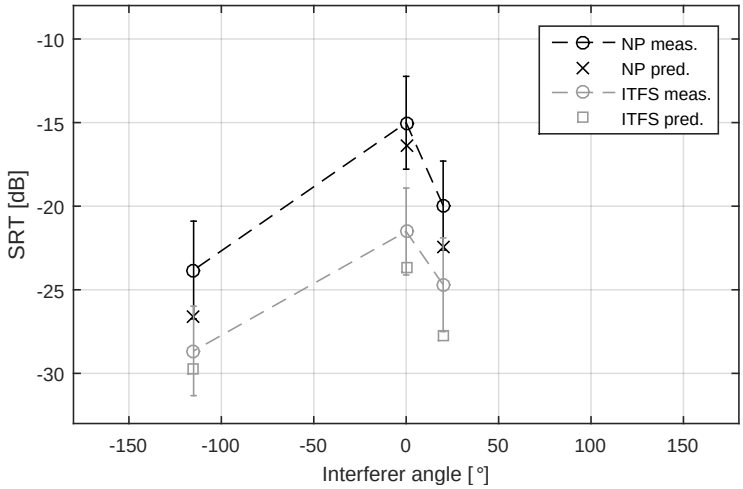


Fig. D.6: The results of conditions 4–6 (NP) and 7–9 (ITFS) of Experiment 3. These conditions are shown together because they all use bottling factory noise for masking. SRTs were estimated from the measured results for each subject individually for each condition. The results, averaged across subjects, are shown as dotted lines with standard deviations. The SRT were predicted with the proposed measure using the reference condition shown in Fig. D.5.

Dantale II sentences and interferer signals of the same length.

Fig. D.5 shows the DBSTOI predictions and the average measured SRTs for the conditions with an SSN interferer (conditions 1–3) together with the measured and predicted SRTs from Experiment 2. It is apparent that ITFS leads to a large advantage in terms of speech intelligibility, as expected. Furthermore, the SRTs in the ITFS-conditions are predicted with an error of less than a standard deviation of the measurements. This indicates that the DBSTOI measure can account for the joint effect of binaural advantage and processing by ITFS. The results of the conditions with bottling factory noise masking are shown in Fig. D.6. The large standard deviations of the measurements indicate that there are large differences between subjects for this masker type. Furthermore, predictions are biased downwards by 2–3 dB. This may be caused by the fact that predictions were made with a reference condition were another type of masker (SSN) was used. However, the relative differences in SRTs between the conditions appear to be rather accurately predicted.

A noteworthy feature of the results relates to binaural advantage. We define binaural advantage in this experiment as the difference in SRT between the S_0N_0-condition and the S_0N_{-115}-condition. Predicted and measured values of binaural advantage for SSN and bottling factory noise with and without ITFS are shown in Table D.3. From this table it can be seen that for both interferer types, the binaural advantage decreases, when the signals are pro-

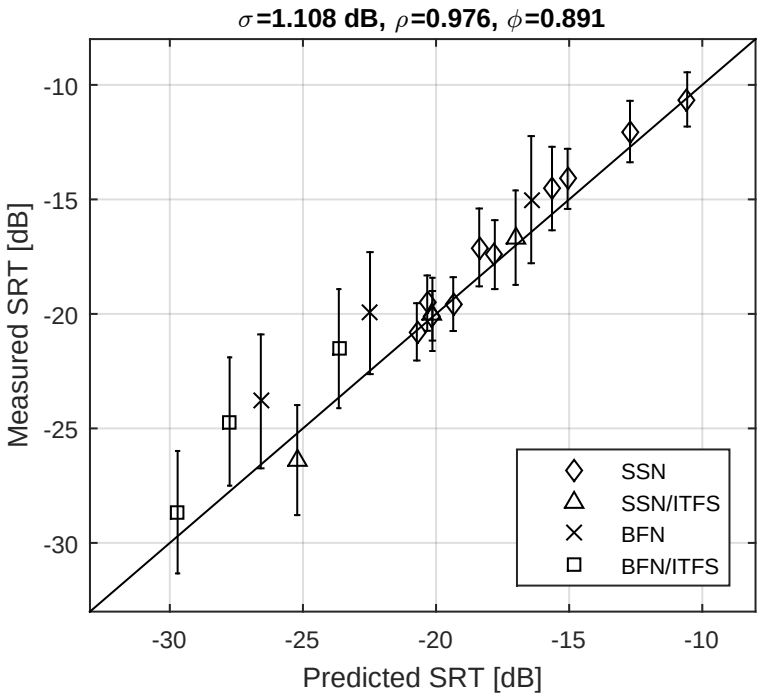


Fig. D.7: Comparison of measured and predicted SRTs from all conditions of Experiment 2 and Experiment 3 with both SSN and bottling factory noise (BFN) interferers. SRTs were estimated separately for each subject for each conditions. The shown SRT values are averaged across subjects with standard deviations shown as error bars. The diagonal line represents perfect predictions. Measures of accuracy in the top of the plot are computed by comparing the measured and predicted SRTs.

cessed with ITFS. A possible explanation of this is that ITFS improves the spectral features of the signal but fails to restore the phase, which is important for binaural advantage. It can be noted that this decrease in binaural advantage is indeed predicted by the DBSTOI measure. The decrease is, however, predicted to be larger than the actual values.

Fig. D.7 compares predictions and measurements for both masker types. An average prediction error of slightly over 1 dB is obtained: an error dominated by the bias of predictions with bottling factory masking.

5.4 Multiple Interferers and Beamforming

Lastly, we consider the results of Experiment 4 (Sec. 3.4), which involves multiple interferers and beamforming. Predictions were made with the DBSTOI measure on the basis of 30 concatenated Dantale II sentences. Fig. D.8 compares measured intelligibility and outputs of DBSTOI. In most of the con-

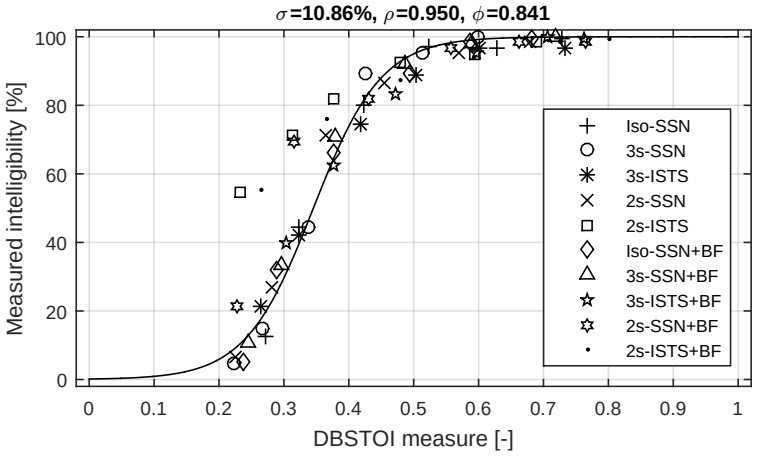


Fig. D.8: Measured intelligibility, averaged across subjects, for the conditions/SNRs of Experiment 4 compared with the predictions of the proposed measure. The shown logistic curve was maximum likelihood fitted to the data. The statistics σ and ρ were computed by comparing the shown data points with the predictions made by the fitted logistic curve. The legend displays the layout of interferers: isotropic (`Iso`), 3 sources in 110°, 180° and −110° (`3s`), 2 sources in 30° and 180° (`2s`). Furthermore it shows the interferer type and whether or not beamforming was included (`BF`).

ditions there appears to be a very strong relationship between the DBSTOI score and the measured results. Especially, the impact of beamforming is well predicted. At low SNRs, there is a discrepancy between predictions made for some of the conditions with two interferers, and the remaining conditions. This may indicate that the DBSTOI measure does not provide fully consistent predictions when making comparisons across different complicated acoustical scenes. If this is the case, the measure should be separately calibrated to acoustically significantly different conditions (e.g. conditions with different numbers of maskers). However, this topic is outside the scope of the present work and has yet to be investigated in depth.

6 Conclusions

We have proposed a binaural speech intelligibility measure based on combining the Short-Time Objective Intelligibility (STOI) measure with an Equalization Cancellation (EC) stage. The proposed measure excels by being capable of predicting the impact of both binaural advantage and non-linear signal processing simultaneously. This makes the measure a potentially powerful tool for the development of signal processing devices which present speech binaurally to a user. The measure outputs ratings of intelligibility on an ar-

bitrary scale, which is useful for comparing e.g. different speech processing algorithms. The measure can be calibrated such as to make direct predictions of Speech Reception Thresholds (SRTs) or of the percentage of correctly understood words. This is useful for predicting the outcome of listening experiments. The accuracy of the measure was investigated by comparing predictions with the results of four listening experiments. The measure was shown to predict the effect of Ideal Time Frequency Segregation (ITFS), with an accuracy similar to that of the original STOI measure. The measure was also shown to predict the effect of binaural advantage, in case of masking by a single point noise source, with an accuracy similar to that of two existing binaural methods. Furthermore, the measure was shown to accurately predict the effect of simultaneous binaural advantage and ITFS. Lastly, the measure was shown to predict well the effect of beamforming, in conditions with multiple interferers. The measure, however, showed some discrepancies when comparing between different conditions with multiple interferers. A detailed investigation of this issue is left for future work. The broad domain in which the measure is applicable calls for further investigation of performance in different conditions, e.g. different types of processing/distortion, different types of interference and different acoustical conditions. Finally, with respect to the particular application of hearing aid signal processing, future work could be directed towards incorporating a hearing loss model into the DBSTOI measure.

A Appendix A

We derive an expression for the expectation of $\mathbf{x}_{q,m}^{\intercal}\mathbf{y}_{q,m}$ under the gaussian random variables, $\Delta\gamma$ and $\Delta\tau$, introduced in the EC stage. We make the assumption that all energy in each third octave band is contained at the center frequency. From (D.9), we obtain:

$$
\begin{aligned}
&E_{\Delta}\left[\mathbf{x}_{q,m}^{\intercal}\mathbf{y}_{q,m}\right] \approx \\
&\quad E_{\Delta}\Big[e^{2\beta+2\Delta\beta}\mathbf{x}_{q,m}^{(l)\intercal}\mathbf{y}_{q,m}^{(l)} + e^{-2\beta-2\Delta\beta}\mathbf{x}_{q,m}^{(r)\intercal}\mathbf{y}_{q,m}^{(r)} \\
&+ \mathbf{x}_{q,m}^{(l)\intercal}\mathbf{y}_{q,m}^{(r)} + \mathbf{x}_{q,m}^{(r)\intercal}\mathbf{y}_{q,m}^{(l)} \\
&- 2(e^{\beta+\Delta\beta}\mathbf{x}_{q,m}^{(l)\intercal} + e^{-\beta-\Delta\beta}\mathbf{x}_{q,m}^{(r)\intercal})\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{y}_{q,m}^{(c)}\right] \\
&- 2\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{x}_{q,m}^{(c)\intercal}\right](e^{\beta+\Delta\beta}\mathbf{y}_{q,m}^{(l)} + e^{-\beta-\Delta\beta}\mathbf{y}_{q,m}^{(r)}) \\
&+ 4\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{x}_{q,m}^{(c)\intercal}\right]\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{y}_{q,m}^{(c)}\right]\Big], \qquad \text{(D.22)}
\end{aligned}
$$

where β is given by (D.13) and:

$$\Delta\beta = \frac{\ln(10)}{20}\Delta\gamma. \tag{D.23}$$

Because the expectation operator is linear we may evaluate the terms of (D.22) independently. Furthermore, since $\Delta\beta$ is zero-mean normally distributed with variance $\sigma^2_{\Delta\beta}$, we have:

$$\begin{aligned}
E_\Delta[e^{\beta+\Delta\beta}] &= e^{\beta}E_\Delta[e^{\Delta\beta}] = e^{\beta}e^{\sigma^2_{\Delta\beta}/2},\\
E_\Delta[e^{-\beta-\Delta\beta}] &= e^{-\beta}E_\Delta[e^{-\Delta\beta}] = e^{-\beta}e^{\sigma^2_{\Delta\beta}/2},\\
E_\Delta[e^{2\beta+2\Delta\beta}] &= e^{2\beta}E_\Delta[e^{2\Delta\beta}] = e^{2\beta}e^{2\sigma^2_{\Delta\beta}},\\
E_\Delta[e^{-2\beta-2\Delta\beta}] &= e^{-2\beta}E_\Delta[e^{-2\Delta\beta}] = e^{-2\beta}e^{2\sigma^2_{\Delta\beta}}.
\end{aligned} \tag{D.24}$$

Using the above allows for computing the expectation of terms 1–4 in (D.22). For terms 5–6, we may make use of the fact that:

$$E_\Delta[f(\Delta\beta)g(\Delta\tau)] = E_\Delta[f(\Delta\beta)]E_\Delta[g(\Delta\tau)],$$

because $\Delta\beta$ and $\Delta\tau$ are statistically independent (where $f, g : \mathbb{C} \to \mathbb{C}$ are any functions). Furthermore, we note that $\mathrm{Re}[ab] = \mathrm{Re}[a]\mathrm{Re}[b] - \mathrm{Im}[a]\mathrm{Im}[b]$ for any $a, b \in \mathbb{C}$. This allows us to write:

$$\begin{aligned}
\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{x}^{(c)}_{q,m}\right] &= \mathrm{Re}\left[e^{-j\omega_q\Delta\tau}\right]\mathrm{Re}\left[e^{-j\omega_q\tau}\mathbf{x}^{(c)}_{q,m}\right]\\
&\quad - \mathrm{Im}\left[e^{-j\omega_q\Delta\tau}\right]\mathrm{Im}\left[e^{-j\omega_q\tau}\mathbf{x}^{(c)}_{q,m}\right].
\end{aligned} \tag{D.25}$$

By (D.24) and by symmetry, respectively, we have:

$$\begin{aligned}
E_\Delta\left[\mathrm{Re}\left[e^{-j\omega_q\Delta\tau}\right]\right] &= e^{-\omega_q^2\sigma^2_{\Delta\tau}/2},\\
E_\Delta\left[\mathrm{Im}\left[e^{-j\omega_q\Delta\tau}\right]\right] &= 0,
\end{aligned} \tag{D.26}$$

which in turn leads to:

$$E_\Delta\left[\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{x}^{(c)}_{q,m}\right]\right] = e^{-\omega_q^2\sigma^2_{\Delta\tau}/2}\mathrm{Re}\left[e^{-j\omega_q\tau}\mathbf{x}^{(c)}_{q,m}\right]. \tag{D.27}$$

To evaluate the last term in (D.22), we note that:

$$\mathrm{Re}\left[a\right]\mathrm{Re}\left[b\right] = \frac{1}{2}\left(\mathrm{Re}\left[ab\right] + \mathrm{Re}\left[a^*b\right]\right),$$

where $a, b \in \mathbb{C}$ and * represents complex conjugation:

$$\begin{aligned}
&E_\Delta\left[4\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{x}^{(c)\mathsf{T}}_{q,m}\right]\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{y}^{(c)}_{q,m}\right]\right] =\\
&2\left(E_\Delta\left[\mathrm{Re}\left[e^{-j2\omega_q(\tau+\Delta\tau)}\mathbf{x}^{(c)\mathsf{T}}_{q,m}\mathbf{y}^{(c)}_{q,m}\right]\right] + \mathrm{Re}\left[\mathbf{x}^{(c)\mathrm{H}}_{q,m}\mathbf{y}^{(c)}_{q,m}\right]\right) =\\
&2\left(e^{-2\omega_q^2\sigma^2_{\Delta\tau}}\mathrm{Re}\left[e^{-j2\omega_q\tau}\mathbf{x}^{(c)\mathsf{T}}_{q,m}\mathbf{y}^{(c)}_{q,m}\right] + \mathrm{Re}\left[\mathbf{x}^{(c)\mathrm{H}}_{q,m}\mathbf{y}^{(c)}_{q,m}\right]\right),
\end{aligned} \tag{D.28}$$

where $(\cdot)^{\mathrm{H}}$ denotes the conjugate transpose. By inserting (D.24), (D.27) and (D.28) into (D.22) (with appropriate substitution of variables), one reaches (D.12) as desired.

Acknowledgment

This work was funded by the Oticon Foundation and the Danish Innovation Foundation. We thank the authors of [37] for making an implementation of their method available.

References

[1] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A method for predicting the intelligibility of noisy and non-linearly enhanced binaural speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China: IEEE, Mar. 2016.

[2] H. Fletcher and J. C. Steinberg, "Articulation testing methods," *Bell System Technical Journal*, vol. 8, no. 4, pp. 806–854, Oct. 1929.

[3] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[4] S. Jørgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acta Acustica United with Acustica*, vol. 101, pp. 1016–1025, 2015.

[5] S. van Wijngaarden, "The speech transmission index after four decades of development," *Acoustics Australia*, vol. 40, no. 2, pp. 134–138, Aug. 2012.

[6] *Sound System Equipment. Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index*, IEC Std.

[7] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for inteligibility prediction of time-frequency weighted noisy speech," *IEEE Tran. on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[8] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[9] C. Ludvigsen, C. Elberling, , and G. Keidser, "Evaluation of a noise reduction method – comparison between observed scores and scores predicted from STI," *Scand. Audiol. Suppl.*, vol. 38, pp. 50–55, 1993.

[10] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3679–3689, Dec. 2004.

[11] J. B. Boldt and D. P. W. Ellis, "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in *Proceedings of the 17th European Signal Processing Conference (EUSIPCO-2009)*. Glasgow, Scotland: EURASIP, Aug. 2009.

[12] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE Tran. on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 430–440, Feb. 2014.

[13] K. D. Kryter, "Methods for the Calculation of and Use of the Articulation Index," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1689–1697, Nov. 1962.

[14] J. B. Allen, "The articulation index is a shannon channel capacity," in *Auditory Signal Processing: Physiology, Psychoacoustics and Models*, D. Pressnitzer, A. de Cheveigne, S. McAdams, and L. Collet, Eds. Springer Verlag, 2004, pp. 314–320.

[15] *Methods for Calculation of the Speech Intelligibility Index*, ANSI Std. S3.5-1997, 1997.

[16] T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355–367, Jan. 1971.

[17] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[18] I. M. Noordhoek and R. Drullmann, "Effect of reducing temporal intensity modulations on sentence intelligibility," *J. Acoust. Soc. Am.*, vol. 101, no. 1, pp. 498–502, Jan. 1997.

[19] V. Hohmann, "The effect of multichannel dynamic compression on speech intelligibility," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1191–1195, Feb. 1995.

[20] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interupted noise," *J. Acoust. Soc. Am.*, vol. 126, no. 6, pp. 3236–3245, Dec. 2009.

[21] F. Dubbelboer and T. Houtgast, "A detailed study on the effects of noise on speech intelligibility," *J. Acoust. Soc. Am.*, vol. 122, no. 5, pp. 2865–2871, Nov. 2007.

[22] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Am.*, vol. 130, no. 3, pp. 1475–1487, Sep. 2011.

[23] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.

[24] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, Dec. 2006.

[25] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, Apr. 2005.

[26] J. Ma and Y. H. P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3387–3405, May 2009.

[27] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, Jul. 2014.

[28] K. Smeds, A. Leijon, F. Wolters, A. Hammarstedt, S. Båsjö, and S. Hertzman, "Comparison of predictive measures of speech recognition after noise reduction processing," *J. Acoust. Soc. Am.*, vol. 136, no. 3, pp. 1363—-1374, Sep. 2014.

[29] A. W. Bronkhorst, "The cocktail party phenomenon: A review on speech intelligibility in multiple-talker conditions," *Acta Acustica United with Acustica*, vol. 86, no. 1, pp. 117–128, Jan. 2000.

[30] N. I. Durlach, "Equalization and cancellation theory of binaural masking-level differences," *J. Acoust. Soc. Am.*, vol. 35, no. 8, pp. 1206–1218, Aug. 1963.

[31] ——, "Binaural signal detection: Equalization and cancellation theory," in *Foundations of Modern Auditory Theory Volume II*, J. V. Tobias, Ed. New York: Academic Press, 1972, pp. 371–462.

[32] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 331–342, Apr. 2006.

[33] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension and evaluation of a binaural speech intelligibility model," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2479–2497, Dec. 2010.

[34] S. J. van Wijngaarden and R. Drullman, "Binaural intelligibility prediction based on the speech transmission index," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4514–4523, Jun. 2008.

[35] R. Wan, N. I. Durlach, and H. S. Colburn, "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," *J. Acoust. Soc. Am.*, vol. 128, no. 6, pp. 3678–3690, Sep. 2010.

[36] M. Lavandier and J. F. Culling, "Prediction of binaural speech intelligibility against noise in rooms," *J. Acoust. Soc. Am.*, vol. 127, no. 1, pp. 387–399, Jan. 2010.

[37] S. Jelfs, J. F. Culling, and M. Lavandier, "Revision and validation of a binaural model for speech intelligibility in noise," *Hearing Research*, vol. 275, no. 1–2, pp. 96–104, May 2011.

[38] M. Lavandier, S. Jelfs, J. F. Culling, A. J. Watkins, A. P. Raimond, and S. J. Makin, "Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources," *J. Acoust. Soc. Am.*, vol. 131, no. 1, pp. 218–231, Jan. 2012.

[39] J. Rennies, T. Brand, and B. Kollmeier, "Prediction of the intelligibility of reverberation on binaural speech intelligibility in noise and in quiet," *J. Acoust. Soc. Am.*, vol. 130, no. 5, pp. 2999–3012, Nov. 2011.

[40] J. Rennies, A. Warzybok, T. Brand, and B. Kollmeier, "Modelling the effects of a single reflection on binaural speech intelligibility," *J. Acoust. Soc. Am.*, vol. 135, no. 3, pp. 1556–1567, Mar. 2014.

[41] J. F. Culling, M. L. Hawley, and R. Y. Litovsky, "The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources," *J. Acoust. Soc. Am.*, vol. 116, no. 2, pp. 1057–1065, Aug. 2004.

[42] ——, "Erratum: The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources [J. Acoust. Soc. Am. 116, 1057 (2004)]," *J. Acoust. Soc. Am.*, vol. 118, no. 1, p. 552, Jul. 2005.

[43] T. Leclère, M. Lavandier, and J. F. Culling, "Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking and binaural de-reverberation," *J. Acoust. Soc. Am.*, vol. 137, no. 6, pp. 3335–3345, Jun. 2015.

[44] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A binaural short time objective intelligibility measure for noisy and enhanced speech," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015.

[45] H. vom Hövel, "Zur bedeutung der übertragungseigenschaften des aussenohrs sowie des binauralen hörsystems bei gestörter sprachübertragung," Ph.D. dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, 1984.

[46] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Matching pursuit for channel selection in coclear implants based on an intelligibility metric," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO-2012)*. Bucharest, Romania: EURASIP, Aug. 2012, pp. 504–508.

[47] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech," *J. Acoust. Soc. Am.*, vol. 130, no. 5, pp. 3013–3027, Nov. 2011.

[48] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.

[49] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *International Journal of Audiology*, vol. 42, no. 1, pp. 10–17, Jan. 2003.

[50] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proccedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, Oct. 2001.

[51] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 123, no. 3, pp. 1673–1682, Mar. 2008.

[52] M. Vestergaard, "The eriksholm cd 01: Speech signals in various acoustical environments," 1998.

[53] E. R. Pedersen and P. M. Juhl, "Speech in noise test based on a ten-alternative forced choice procedure," in *Joint Baltic-Nordic Acoustics Meeting (BNAM)*, Odense, Denmark, Jun. 2012.

[54] ——, "User-operated speech in noise test: Implementation and comparison with a traditional test," *International Journal of Audiology*, vol. 53, no. 5, pp. 336–344, May 2014.

[55] I. Holube, S. Fredelake, M. Vlaming, and B. Kollmeier, "Development and analysis of an international speech test signal (ISTS)," *International Journal of Audiology*, vol. 49, no. 12, pp. 891–903, Dec. 2010.

[56] T. Brand and B. Kollmeier, "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," *J. Acoust. Soc. Am.*, vol. 111, no. 6, pp. 2801–2810, Jun. 2002.

[57] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, Jun. 1938.

Paper E

A Non-Intrusive Short-Time Objective Intelligibility Measure

Asger Heidemann Andersen
Jan Mark de Haan
Zheng-Hua Tan
Jesper Jensen

The paper has been presented at the
42st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, United States, pp. 5085-5089, 2017.

© 2017 IEEE
The layout has been revised.

Abstract

We propose a non-intrusive intelligibility measure for noisy and non-linearly processed speech, i.e. a measure which can predict intelligibility from a degraded speech signal without requiring a clean reference signal. The proposed measure is based on the Short-Time Objective Intelligibility (STOI) measure. In particular, the non-intrusive STOI measure estimates clean signal amplitude envelopes from the degraded signal. Subsequently, the STOI measure is evaluated by use of the envelopes of the degraded signal and the estimated clean envelopes. The performance of the proposed measure is evaluated on a dataset including speech in different noise types, processed with binary masks. The measure is shown to predict intelligibility well in all tested conditions, with the exception of those including a single competing speaker. While the measure does not perform as well as the original (intrusive) STOI measure, it is shown to outperform existing non-intrusive measures.

1 Introduction

In recent years, Speech Intelligibility Prediction (SIP) has been investigated with great interest due to its potential as a tool in optimizing speech intelligibility across a wide range of applications, including e.g. architectural acoustics [1], telecommunications [2], and hearing aid signal processing [3–5]. Much of the recent work in the field is based on the classical methods: the Speech Intelligibility Index (SII) [6, 7] and the Speech Transmission Index (STI) [8, 9]. This has led to methods such as the Extended SII (ESII) [10, 11], the Coherence SII (CSII) [12] and the Binaural Speech Intelligibility Measure (BSIM) [13, 14]. Recently, the physiologically founded multi-resolution sEPSM (mr-sEPSM) [15] has received attention for its ability to predict intelligibility of speech in reverberation and modulated noise. Another recent method, the Short-Time Objective Intelligibility (STOI) measure [16], has become popular within the signal processing community because of its simplicity and proven ability to predict the impact of various speech processing algorithms [2, 5, 17]. Several variations of the STOI measure with specialized properties have been proposed [3, 18, 19].

The mentioned methods require access to a clean reference signal in addition to either the masker signal or the degraded speech signal. These are referred to as intrusive methods, because of their dependence on a clean reference signal. In some situations, intrusive methods cannot be applied because the clean reference signal is unknown or poorly defined, e.g. when attempting to predict the intelligibility of an unknown speech signal on a signal processing device in realtime.

The above concern has led to research into non-intrusive SIP. One such method is the Speech to Reverberation Modulation energy Ratio (SRMR) [20]

which aims to predict the intelligibility of reverberated speech from the ratio between low and high modulation frequency energy. The SRMR has been shown to outperform a number of existing measures [20]. While originally formulated to predict the intelligibility of reverberant signals, the authors have later used the measure successfully to predict the intelligibility of noisy and processed signals [5]. A similar measure, the average modulation-spectrum area (ModA) [21], aims to predict the intelligibility of reverberated speech from the area of the modulation spectrum. This measure has been shown to compare favorably to other non-intrusive methods across a range of conditions spanning reverberation, additive noise, and distortion [5].

Another means to obtain non-intrusive SIP methods, is to estimate the output of existing intrusive methods, without using a clean reference signal. This can be done using machine learning, or by using noise reduction to estimate the clean signal from the degraded one. For instance, [22] uses a twin Hidden Markov Model (HMM) to estimate the STOI measure, while [23] uses tree based regression to predict both the STOI and PESQ [24] measures. A semi-non-intrusive method for hearing aids, using beamforming to estimate the clean signal, is proposed in [4] .

In the present paper we propose a fully non-intrusive version of the STOI measure. The proposed measure estimates envelopes of the clean reference signal from the degraded signal by use of a statistical clean speech model. The measure requires training with clean speech, but does not require training with particular interferer- or processing types. The remainder of the paper progresses as follows: Sec. 2 describes the proposed measure, Sec. 3 evaluates the measure and compares it to a number of existing intelligibility measures, and Sec. 4 concludes upon the presented findings.

2 The NI-STOI Measure

We describe the proposed Non-Intrusive STOI (NI-STOI) measure, which is similar to the original (intrusive) STOI measure [16]. The STOI measure assumes intelligibility to be related to the correlation between clean and degraded 1/3-octave band amplitude envelopes. However, as we do not assume a clean reference signal to be available, we estimate the clean speech envelopes from the degraded speech envelopes. This is done by use of a statistical model of clean speech. An overview of the NI-STOI measure is given in Fig. E.1.

2.1 Generating a Clean Speech Model

To distinguish between speech and noise/distortion, we generate a modulation domain model of clean speech. This model is generated on the basis of a

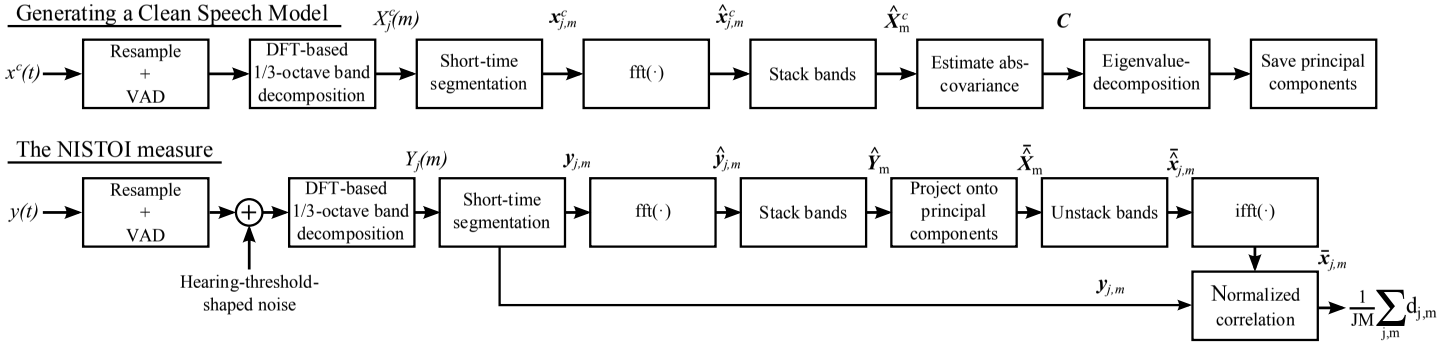


Fig. E.1: A block diagram of the proposed non-intrusive intelligibility measure. The top illustrates how a clean speech model is generated from a sequence of clean speech, $x_c(t)$. The bottom illustrates how the intelligibility measure is evaluated for a degraded signal, $y(t)$.

long clean speech signal, $x^c(t)$, i.e. long enough to be considered representative of speech in general. Silent parts of the signal are removed with a Voice Activity Detector (VAD), and the signal is resampled to 10 kHz, as for the original STOI measure [16]. The resulting signal is Time-Frequency (TF) decomposed with a short time Discrete Fourier Transformation (DFT) as specified in [16]. Let $\hat{x}^c(k,m) \in \mathbb{C}$ denote the kth DFT-coefficient of the mth window.

We then extract $J = 15$ 1/3-octave band envelopes from the TF-decomposed signal as follows [16]:

$$X_j^c(m) = \sqrt{\sum_{k_1(j)}^{k_2(j)} |\hat{x}^c(k,m)|^2}, \tag{E.1}$$

where $k_1(j)$ and $k_2(j)$ are the lower and upper bounds of the jth 1/3-octave band. The resulting envelope samples are arranged in vectors of $N = 30$ samples:

$$\mathbf{x}_{j,m}^c = \left[X_j^c(m-N+1), \ldots, X_j^c(m)\right]^T, \tag{E.2}$$

which are normalized to have zero mean and unit norm:

$$\tilde{\mathbf{x}}_{j,m}^c = \frac{\mathbf{x}_{j,m}^c - \mathbf{1}\mu_{\mathbf{x}_{j,m}^c}}{||\mathbf{x}_{j,m}^c||}, \tag{E.3}$$

where $\mu_{(\cdot)}$ denotes the mean of entries in a vector and $\mathbf{1}$ is a vector of ones. Let $\hat{\mathbf{x}}_{j,m}^c$ denote the DFT of $\tilde{\mathbf{x}}_{j,m}^c$, i.e. the modulation domain representation of the signal. We then stack modulation domain representations for all frequency bands into one vector:

$$\hat{\mathbf{X}}_m^c = \left[\hat{\mathbf{x}}_{1,m}^{c\,T}, \ldots, \hat{\mathbf{x}}_{J,m}^{c\,T}\right]^T \in \mathbb{R}^{JN\times 1}. \tag{E.4}$$

The transition from (E.2) to (E.4) is illustrated on Fig. E.1 by the blocks "fft(·)" and "Stack bands". We use the resulting vectors to estimate an amplitude covariance matrix for all modulation frequencies across all frequency bands:

$$\mathbf{C} = \frac{1}{M}\sum_{m=N}^{M+N-1} |\hat{\mathbf{X}}_m^c|\,|\hat{\mathbf{X}}_m^c|^T, \tag{E.5}$$

where M is the number of frames, and $|\cdot|$ denotes the absolute value which is evaluated on an entry-wise basis for vectors. As we show in Sec 3, matrix $\mathbf{C}$ can be approximated well by a low rank matrix. This property can be used to distinguish between speech and non-speech components of degraded speech envelopes. We compute the eigenvalue decomposition of $\mathbf{C}$, resulting

in a descending sequence of eigenvalues, $\lambda_1, \lambda_2, \ldots, \lambda_{JN}$, and corresponding eigenvectors, $v_1, v_2, \ldots, v_{JN}$. In the following section we use the principal components, $v_1, \ldots, v_K$, with $1 \leq K \leq JN$, to estimate the envelopes of the unknown clean reference signal.

2.2 Computing the NI-STOI Measure

We now describe the computation of the proposed NI-STOI measure. This is done as for the original STOI measure [16], except that the clean envelope samples are estimated from the degraded ones, because only the degraded signal, $y(t)$, is assumed known. Silent regions of the signals are removed, by use of the same VAD as for the original STOI measure. While this *does* make use of the clean reference signal, $x(t)$ it ensures comparability with the original STOI measure. Then, the degraded signal is resampled to 10 kHz. At this stage we add a faint noise signal, shaped such that the energy in each 1/3-octave band corresponds to the average hearing threshold in quiet [25] (similar to what is done in e.g. [13]). This has shown necessary in conditions where aggressive speech processing renders the presented signal almost inaudible. The noise has little impact on predictions at normal speech levels. A TF decomposition, carried out as described in Sec. 2.1, results in DFT coefficients of the degraded signal, $\hat{y}(k, m)$. Using these, we define envelope samples, $Y_j(m)$, similar to (E.1), normalized envelope vectors, $\tilde{\mathbf{y}}_{j,m}$, similar to (E.3), and modulation domain vectors, $\hat{\mathbf{Y}}_m$, similar to (E.4). We then construct an estimate of the corresponding clean signal modulation vector, $\hat{\mathbf{X}}_m$, by assuming: 1) the phase of $\hat{\mathbf{X}}_m$ is the same as the phase of $\hat{\mathbf{Y}}_m$, and 2) the magnitude of $\hat{\mathbf{X}}_m$ can be approximated by projecting the magnitude of $\hat{\mathbf{Y}}_m$ into the space spanned by the K clean signal principal components, $v_1, \ldots, v_K$, found in Sec. 2.1. These assumptions lead to the following estimate of $\hat{\mathbf{X}}_m$:

$$\bar{\hat{\mathbf{X}}}_m = e^{j\angle \hat{\mathbf{Y}}_m} \odot \sum_{k=1}^{K} v_k v_k^T |\hat{\mathbf{Y}}_m|, \tag{E.6}$$

where $e^{j\angle \hat{\mathbf{Y}}_m}$ is a vector in which all entries have the same phase as $\hat{\mathbf{Y}}_m$, but unit magnitude, while $\odot$ denotes entry-wise multiplication (Hadamard product). The resulting estimate is split into J vectors, $\bar{\hat{\mathbf{x}}}_{1,m}, \ldots, \bar{\hat{\mathbf{x}}}_{J,m}$, of length N, corresponding to the inverse of the operation described in (E.4). By computing the DFT of these vectors, we obtain estimates, $\bar{\mathbf{x}}_{1,m}, \ldots, \bar{\mathbf{x}}_{J,m}$, of the clean signal envelopes. From this point, we can compute the (intrusive) STOI measure, using the estimated clean speech envelopes in place of the true ones. To do this, we compute the correlation between the (estimated) clean and

degraded envelopes [16][1]:

$$d_{j,m} = \frac{\left(\bar{\mathbf{x}}_{j,m} - \mathbf{1}\mu_{\bar{\mathbf{x}}_{j,m}}\right)^T \left(\mathbf{y}_{j,m} - \mathbf{1}\mu_{\mathbf{y}_{j,m}}\right)}{||\bar{\mathbf{x}}_{j,m} - \mathbf{1}\mu_{\bar{\mathbf{x}}_{j,m}}||||\mathbf{y}_{j,m} - \mathbf{1}\mu_{\mathbf{y}_{j,m}}||}. \tag{E.7}$$

The NI-STOI measure is then computed, as described in [16], as the average normalized correlation between clean and degraded envelopes:

$$\text{NI-STOI} = \frac{1}{JM} \sum_{j,m} d_{j,m}. \tag{E.8}$$

In order to carry out direct predictions of intelligibility, in terms of a percentage of correctly answered words, the output of the NI-STOI measure is transformed with a logistic function [16]:

$$\bar{s}(x) = \frac{100\%}{1 + e^{ax+b}}, \tag{E.9}$$

where x is the input NI-STOI measure and $\bar{s}$ is the estimated intelligibility in percent. The coefficients a and b are fitted to available data by maximum likelihood (as described in [27]).

3 Results and Discussion

In this section we first show results from the training of a number of clean speech models. We then evaluate the proposed non-intrusive intelligibility measure by using it to predict the results of a listening experiment.

3.1 Clean Speech Models

To investigate the dependence on clean speech material, we train clean speech models as described in Sec. 2.1, using three different sources of clean speech: 1) all the sentences from the Dantale II corpus [28], 2) all sentences with female speakers from the TIMIT training corpus [29], 3) all sentences, male and female speakers, from the TIMIT training corpus.

Fig. E.2 shows the cumulative sum of descending eigenvalues for the three models. For all three models, the majority of the energy in the modulation magnitude spectra can be accounted for by a single principal component. This can be seen as an indication that the modulation domain representation

[1]The original STOI measure includes a clipping stage which serves to limit the extent to which a single frame can be detrimental to overall predicted intelligibility, in cases where there is very little, or negative, correlation between the clean and degraded envelopes. We have chosen not to include this stage in the NI-STOI measure. The removal of the clipping mechanism has previously been shown not to decrease performance markedly [3, 26].

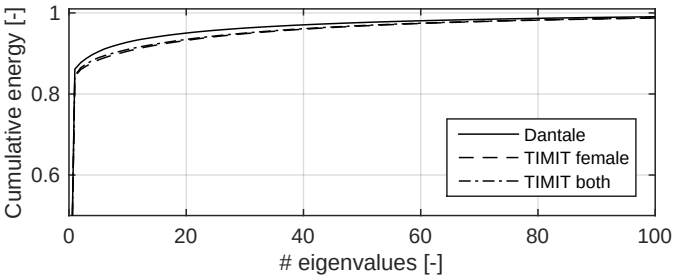


Fig. E.2: The normalized cumulative sum of eigenvalues of **C** when generated with each of the three different clean speech corpora.

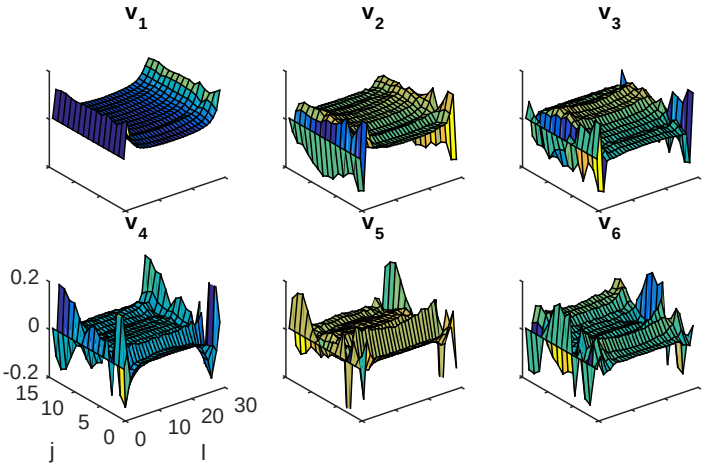


Fig. E.3: The first six principal components, obtained by training with clean speech from the Dantale II corpus. The axes, as shown on the lower left plot, are identical for all the plots. The j-axis denotes the 15 1/3-octave bands, while the l-axis denotes the 30 modulation frequency bins resulting from the DFT of an envelope vector. Note that the plots are symmetric on the l-axis.

is a strong starting point for low dimensional representations of speech. Contrarily, it should be noted that this representation includes neither the phase of the speech signal nor the phase of the envelopes, and therefore cannot be used to reconstruct the original speech signal. Fig. E.3 shows the first six principal components, obtained by training with the Dantale II speech material. Here, the first component, v_1, is most important, as it codes for the majority of the modulation energy. The shape of this component indicates that most of the modulation energy is contained at low modulation frequencies (i.e. less than 10 Hz). This fits well with the rationale of the SRMR measure which considers low frequency modulations to be carriers of speech information.

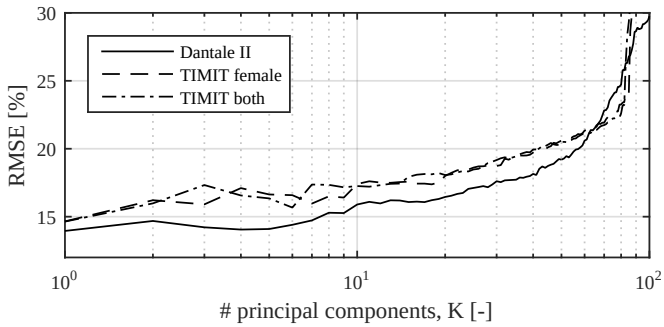


Fig. E.4: Prediction performance in terms of RMSE, of NI-STOI, vs., K, for each of the three clean speech models. The conditions with café noise were excluded in this analysis. See the text for details.

3.2 Experimental Data

To evaluate the performance of the proposed measure, we use it on a set of data [30] which was also used for evaluating the original STOI measure [16]. Intelligibility was measured for 15 normal hearing subjects, using the Dantale II sentence material [28]. Four types of noise was used: bottling factory hall noise, café noise, car noise, and Speech Shaped Noise (SSN), each presented at three different Signal to Noise Ratios (SNRs). The noisy signals were processed with Ideal Binary Masks (IBMs) and Target Binary Masks (TBMs) at eight different Relative Criterion (RC) values[2] [30]. Two sentences were presented with each combination of the above for a total of 15 subjects $\times$ 7 noise/mask combinations $\times$ 8 RC values $\times$ 3 SNRs $\times$ 2 repetitions $=$ 5040 sentences. The dataset is described in detail in [30].

3.3 Predictions

We first consider the overall performance of the NI-STOI measure, and its dependency on the number of principal components, K, and the clean speech material used for training. The conditions with café noise are not included in this analysis, for reasons which are discussed later. Fig. E.4 shows the Root-Mean-Square Error (RMSE) of predictions versus K. The best performance is obtained with the Dantale II speech model. This indicates that, in spite of the simplicity of the applied speech model, some degree of talker specific modeling can occur. Contrarily, the TIMIT-based speech models perform about equally well. This suggests that the models do not capture gender specific effects. This may be due to the absence of pitch information in the speech representation. With regards to the number of principal components, K, there

[2]The RC value is an algorithm parameter which determines the density of the computed binary mask. See [30] for details.

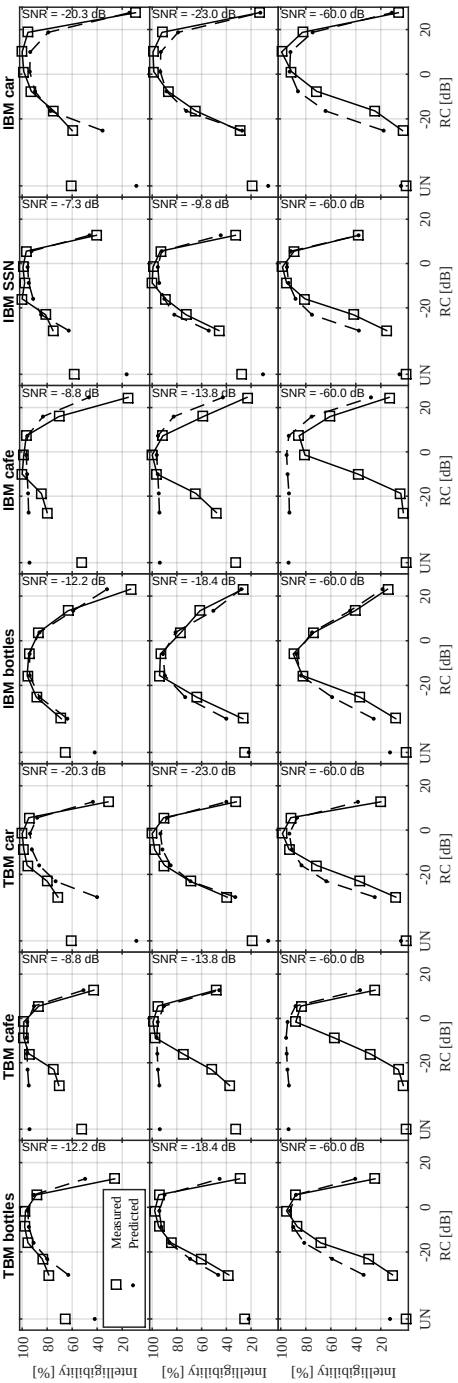


Fig. E.5: NI-STOI predictions with $K = 1$, compared with measured results. The clean speech model based on Dantale II sentences was used, and the logistic mapping function has been fitted to all measured data except the conditions with café noise. Columns correspond to different noise- and processing types, while rows correspond to different input SNRs, decreasing downwards. The horizontal axis shows the RC-value of the processing algorithm. A low RC corresponds to mild processing while high RC corresponds to heavy processing. Unprocessed conditions are denoted "UN".

	Pearson correlation	RMSE	Kendall's τ	
STOI [16]	0.958	9.5%	0.824	÷ café
NI-STOI (Dantale II)	0.907	13.9%	0.777	
NI-STOI (TIMIT F)	0.897	14.6%	0.764	
NI-STOI (TIMIT M+F)	0.897	14.6%	0.768	
SRMR [20]	0.311	42.0%	0.207	
SRMR-norm [31]	0.550	31.6%	0.388	
STOI [16]	0.959	9.4%	0.822	All conds.
NI-STOI (Dantale II)	0.711	25.2%	0.529	
NI-STOI (TIMIT F)	0.704	25.4%	0.516	
NI-STOI (TIMIT M+F)	0.702	25.5%	0.513	
SRMR [20]	0.237	45.2%	0.036	
SRMR-norm [31]	0.394	38.6%	0.156	

Table E.1: Comparison of intelligibility measures with and without café noise. In both cases, the logistic mapping function was fitted without café noise. The NI-STOI measure was used with $K = 1$.

is a clear tendency that more components lead to poorer performance. While performance is relatively constant for K between one and ten, the best performance is obtained for $K = 1$. This corresponds well with the observation, supported by Fig. E.2, that most of the modulation energy is captured by a single principal component. Adding more components adds rather little clean speech information, but may let more noise and distortion into the clean envelope estimate.

Fig. E.5 shows NI-STOI predictions for the individual conditions. Again, the logistic mapping function was fitted without the café noise conditions. The figure indicates a good overall fit between predictions and measurements. Large deviations are seen for the conditions with café noise at low RC-values (corresponding to little or no processing), where intelligibility is predicted to be very high, while in fact it is rather low. It should be noted that the café noise consists mainly of a single interfering female talker. Since the NI-STOI measure is non-intrusive, it has no means of determining which speaker is the target one (assuming that the clean speech model is not talker specific). Therefore, one can argue that any non-intrusive SIP method is bound to fail in such a condition, unless supplied with additional information about which speaker is the target one. This also explains why the overall quality of the logistic mapping can be increased by excluding the café noise conditions during fitting.

In Table E.1, we evaluate the proposed measure against a number of existing measures. Being intrusive, the STOI measure outperforms the NI-STOI measure in all cases. However, when excluding the café noise conditions, the NI-STOI performance comes somewhat close to that of the STOI measure. There are no major differences between the results for the three different clean speech models. We also compare to two variations of the SRMR

measure which the authors have kindly made available to the public: 1) the original SRMR measure [20], and 2) a later version of the measure which has been improved with the aim of lowering the output variability [31]. While these measures are mainly aimed at predicting intelligibility of reverberated speech, they have been successfully applied for noisy and processed speech [5]. As shown in Table E.1, the improved SRMR measure outperforms the original one, especially when the café noise conditions are excluded. However, the NI-STOI measure also consistently outperforms both measures. The higher performance of the NI-STOI measure is especially pronounced in the absence of the café noise conditions. This result should, however, be viewed in the light of the fact that the NI-STOI measure is trained with clean speech material, while the SRMR measure is not trained or fitted in any manner (except for the logistic mapping, (E.9)).

4 Conclusions

We have proposed a non-intrusive intelligibility measure based on the Short-Time Objective Intelligibility (STOI) measure. Similar to the original STOI measure, the proposed measure is aimed at predicting the intelligibility of noisy and non-linearly processed speech. The model estimates unknown clean speech envelopes from degraded envelopes, by use of a clean speech model. The performance of the proposed measure was evaluated with a dataset consisting of speech in different types of noise processed with binary masks. This indicated that the proposed measure performs better than an existing non-intrusive measure, but not as good as the original (intrusive) STOI measure.

References

[1] S. van Wijngaarden, "The speech transmission index after four decades of development," *Acoustics Australia*, vol. 40, no. 2, pp. 134–138, Aug. 2012.

[2] S. Jørgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acta Acustica United with Acustica*, vol. 101, pp. 1016–1025, 2015.

[3] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Predicting the intelligibility of noisy and non-linearly processed binaural speech," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 1908–1920, 2016.

[4] C. Sørensen, J. B. Boldt, F. Gran, and M. G. Christensen, "Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids," in *The European Signal Processing Conference (EUSIPCO)*, Budapest, Hungary, Aug. 2016, pp. 1358–1362, EURASIP.

[5] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[6] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[7] ANSI Std. S3.5-1997, "Methods for calculation of the speech intelligibility index," 1997.

[8] T. Houtgast and H. J. M. Steeneken, "Evaluation of speech transmission channels by using artificial signals," *Acustica*, vol. 25, no. 6, pp. 355–367, Jan. 1971.

[9] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[10] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.

[11] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, Dec. 2006.

[12] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, Apr. 2005.

[13] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 331–342, Apr. 2006.

[14] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension and evaluation of a binaural speech intelligibility model," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2479–2497, Dec. 2010.

[15] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 134, no. 1, pp. 436–446, July 2013.

[16] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for inteligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.

[17] K. Smeds, A. Leijon, F. Wolters, A. Hammarstedt, S. Båsjö, and S. Hertzman, "Comparison of predictive measures of speech recognition after noise reduction processing," *J. Acoust. Soc. Am.*, vol. 136, no. 3, pp. 1363—-1374, Sept. 2014.

[18] L. Lightburn and M. Brookes, "A weighted STOI intelligibility metric based on mutual information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5365–5369, IEEE.

[19] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[20] T. H. Falk, C. Zheng, and W.-Y. Chan, "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1766–1774, Sept. 2010.

[21] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomedical Signal Processing and Control*, vol. 8, pp. 311–314, Dec. 2013.

[22] M. Karbasi, A. H. Abdelaziz, and D. Kolossa, "Twin HMM-based non-intrusive speech intelligibility prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Sept. 2016, pp. 624–628, IEEE.

[23] D. Sharma, Y. Wang, and P. A. Naylor, "A data-driven non-intrusive measure of speech quality and intelligibility," *Speech Communication*, vol. 80, pp. 84–94, Apr. 2016.

[24] "P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[25] B. C. Moore, *An introduction to the psychology of hearing*, Brill, sixth edition, 2013.

[26] Cees H. Taal, R. C. Hendriks, and R. Heusdens, "Matching pursuit for channel selection in coclear implants based on an intelligibility metric,"

in *The European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, Aug. 2012, pp. 504–508, EURASIP.

[27] T. Brand and B. Kollmeier, "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," *J. Acoust. Soc. Am.*, vol. 111, no. 6, pp. 2801–2810, June 2002.

[28] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *International Journal of Audiology*, vol. 42, no. 1, pp. 10–17, Jan. 2003.

[29] DARPA, "TIMIT, acoustic-phonetic continuous speech corpus," .

[30] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, Sept. 2009.

[31] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan les Pins, France, Sept. 2014, pp. 55–59, IEEE.

Paper F

On the use of Band Importance Weighting in the Short-Time Objective Intelligibility Measure

Asger Heidemann Andersen
Jan Mark de Haan
Zheng-Hua Tan
Jesper Jensen

The paper has been presented at
INTERSPEECH, pp. 2963–2967, Stockholm, Sweden, 2017.

© 2017 ISCA
The layout has been revised.

Abstract

Speech intelligibility prediction methods are popular tools within the speech processing community for objective evaluation of speech intelligibility of e.g. enhanced speech. The Short-Time Objective Intelligibility (STOI) measure has become highly used due to its simplicity and high prediction accuracy. In this paper we investigate the use of Band Importance Functions (BIFs) in the STOI measure, i.e. of unequally weighting the contribution of speech information from each frequency band. We do so by fitting BIFs to several datasets of measured intelligibility, and cross evaluating the prediction performance. Our findings indicate that it is possible to improve prediction performance in specific situations. However, it has not been possible to find BIFs which systematically improve prediction performance beyond the data used for fitting. In other words, we find no evidence that the performance of the STOI measure can be improved considerably by extending it with a non-uniform BIF.

1 Introduction

Speech Intelligibility Prediction (SIP) methods are increasingly being used by the speech processing community in lieu of time consuming and expensive listening experiments. Such methods can provide quick and inexpensive estimates of speech intelligibility in conditions where speech is subjected to e.g. additive noise, reverberation, distortion or speech enhancement. An early SIP method is the Articulation Index (AI) [1] which was proposed for the purpose of evaluating the intelligibility of speech transmitted via telephone. A more recent, improved and standardized, version of the AI is known as the Speech Intelligibility Index (SII) [2]. Further modifications of the SII have been proposed with aims of handling e.g. fluctuating masker signals [3, 4], non-linearly distorted speech [5], and binaural signals [6, 7]. More recently, the multi-resolution sEPSM (mr-sEPSM) has received attention for its physiological basis and its ability to predict intelligibility accurately across a wide range of conditions including reverberation, fluctuating maskers, and noise suppression [8]. The Short-Time Objective Intelligibility (STOI) [9] measure has recently gained popularity in the speech processing community. While the measure is simple and easy to use, it has also proven to predict intelligibility accurately in many conditions including e.g. additive noise, speech enhancement [9, 10], distortion from transmission via telephone [11], and hearing impairment [12]. Several variations of the STOI measure with various purposes and properties have recently been proposed [13–16].

All of the above mentioned methods are roughly characterized by the same procedure: 1) split the involved speech signal into narrow frequency bands with a filterbank, thus mimicking the frequency selectivity of the basilar membrane, 2) estimate the amount of speech information conveyed in

each frequency band, and 3) sum the information from all frequency bands, using some relative weighting that reflects how speech information is distributed across frequency. The frequency weighting function used in the third step is often termed a Band Importance Function (BIF). A BIF for the AI is determined in [1] by use of a graphical procedure, based on measured intelligibility of Highpass (HP) and Lowpass (LP) filtered noisy speech. Such BIFs are also used in the more recent SII [2]. The use of these has since spread to other SIP methods which are based on the SII [3–7]. The advent of modern computing has allowed fitting of BIFs, such as to maximize prediction accuracy for particular datsets of measured intelligibility [17]. Lastly, some authors have proposed SIP methods which use signal dependent BIFs, which are computed such as to reflect the instantaneous information distribution of speech across frequency [14, 18].

The STOI measure distances itself from other measures by being designed with a strong focus on simplicity, and therefore does not include any BIFs [9]. Instead, the STOI measure uniformly averages contributions from 15 one-third octave bands. The designers of the STOI measure [9] made this decision purely with the aim of simplicity, and do not report the effect of this decision (with the exception of noting that the resulting measure has a high performance, in spite of the uniform BIF). However, given the importance of BIFs assumed by other SIP methods, it appears likely that the performance of the STOI measure can be improved by extending it with a suitable BIF.

In this paper we investigate the effect of extending the STOI measure with fitted BIFs. In Sec. 2 we describe the STOI measure, including the modification of including BIFs, and following a similar approach given in [17], we describe how BIFs are fitted such as to minimize the prediction error for datasets of measured intelligibility. In Sec. 3 we describe the two datasets of measured intelligibility which we use for fitting BIFs. These datasets are further divided into different subsets. In Sec. 4 we investigate fitted BIFs for the different subsets of measured intelligibility. Sec. 5 concludes upon our findings.

2 Methods

In this Section we outline the concepts we apply in investigating the use of BIFs together with the STOI measure.

2.1 The STOI Measure

The STOI measure estimates the intelligibility of a degraded speech signal, $y(t)$, by comparing it to a clean reference signal, $x(t)$. Both signals are resampled to 10 kHz and silent regions are removed by use of an ideal

Voice Activity Detector (VAD) [9]. The signals are Time-Frequency (TF) decomposed by use of a short time Discrete Fourier Transformation (DFT) (see details in [9]). Let the degraded signal DFT coefficient of the kth frequency bin and the mth frame be denoted $\hat{y}(k,m)$, and the corresponding clean signal DFT coefficient be denoted by $\hat{x}(k,m)$. Envelopes for each of $J = 15$ one-third octave bands are extracted from the DFT coefficients [9]:

$$X_j(m) = \sqrt{\sum_{k=k_1(j)}^{k_2(j)} |\hat{x}(k,m)|^2}, \tag{F.1}$$

where $k_1(j)$ and $k_2(j)$ denotes, respectively, the lower and upper bounds of the jth one-third octave band. The one-third octave bands have center frequencies from 150 Hz and upwards in one-third octave steps. Corresponding envelope samples, $Y_j(m)$, are defined for the degraded signal. The resulting envelope samples are arranged in vectors of $N = 30$ samples [9]:

$$\mathbf{x}_{j,m} = \left[X_j(m-N+1), \ldots, X_j(m)\right]^T. \tag{F.2}$$

Corresponding vectors, $\mathbf{y}_{j,m}$ are defined for the degraded signal. We define a normalized and clipped version of $\mathbf{y}_{j,m}$, such as to minimize the sensitivity of the method to severely degraded TF-units [9]:

$$\bar{\mathbf{y}}_{j,m}(n) = \min\left(\frac{||\mathbf{x}_{j,m}||}{||\mathbf{y}_{j,m}||}\mathbf{y}_{j,m}(n), (1+10^{-\beta/20\text{dB}})\mathbf{x}_{j,m}(n)\right), \tag{F.3}$$

for $n = 1, \ldots, N$, where $\beta = 15$ dB is a lower bound on signal-to-distortion-ratio [9]. The resulting short-time envelope vectors, $\mathbf{x}_{j,m}$ and $\bar{\mathbf{y}}_{j,m}$ are used to define intermediate correlation coefficients [9]:

$$d_{j,m} = \frac{\left(\mathbf{x}_{j,m} - \mathbf{1}\mu_{\mathbf{x}_{j,m}}\right)^T \left(\bar{\mathbf{y}}_{j,m} - \mathbf{1}\mu_{\bar{\mathbf{y}}_{j,m}}\right)}{||\mathbf{x}_{j,m} - \mathbf{1}\mu_{\mathbf{x}_{j,m}}||\ ||\bar{\mathbf{y}}_{j,m} - \mathbf{1}\mu_{\bar{\mathbf{y}}_{j,m}}||}, \tag{F.4}$$

where $\mathbf{1}$ is a vector of ones, and $\mu_{(\cdot)}$ denotes the sample mean of a vector. The STOI measure is then obtained as the average of $d_{j,m}$ across all values of j and m [9]. This implies a uniform weighting (BIF) for all one-third octave bands j. In this paper, to allow for different BIFs, we instead define bandwise average correlations:

$$\tilde{d}_j = \frac{1}{M}\sum_m d_{j,m}, \tag{F.5}$$

where M is the number of time frames. These are averaged with the BIF $\mathbf{w} = [w_1, \ldots, w_J]^T$, to obtain the final frequency weighted STOI score:

$$s = \sum_{j=1}^{J} w_j \tilde{d}_j, \tag{F.6}$$

where $w_j \geq 0$ for $j = 1, \ldots, J$ and $\sum_{j=1}^{J} w_j = 1$.

The resulting STOI score is a number in the range from 0 to 1, where a higher STOI score indicates higher intelligibility (e.g. percentage of words understood correctly). In order to transform the STOI score into a direct estimate of intelligibility in %, a logistic mapping is applied [9]:

$$f(s;\, a,\, b) = \frac{100\%}{1 + \exp(as + b)}, \tag{F.7}$$

where a and b are fitted such as to maximize prediction accuracy on a well-defined dataset of measured intelligibility.

2.2 Fitting of Band Importance Functions

We now turn to the determination of the BIF, $\mathbf{w}$. We determine this, such as to minimize the prediction error in terms of Root-Mean-Square Error (RMSE). This is heavily inspired by the approach taken in [17] (which fits RMSE optimal weights for the SII). Specifically, we assume that speech intelligibility has been measured in L conditions (e.g. different types of reverberation, distortion or processing at different Signal to Noise Ratios (SNRs)), and is given by $p(l)$, $l = 1 \ldots, L$, where $0\% \leq p(l) \leq 100\%$ is the average fraction of correctly repeated words. We furthermore assume that samples of clean and degraded speech are available for each condition, such that we may compute bandwise average correlations, $\tilde{d}_j(1), \ldots, \tilde{d}_j(L)$, with $j = 1, \ldots, J$, for each condition, using (F.5). For a given BIF, $\mathbf{w}$, we can compute a weighted STOI score for each condition, by (F.6). We can further transform this score into a direct prediction of intelligibility by (F.7). The RMSE of this prediction can be written as:

$$\mathrm{RMSE}(\mathbf{w},\, a,\, b) = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(p(l) - f\left(\sum_{j=1}^{J} w_j \tilde{d}_j(l);\, a,\, b\right)\right)^2}. \tag{F.8}$$

We jointly determine a, b and $\mathbf{w}$ such as to minimize the RMSE, as given by (F.8):

$$\begin{aligned} &\underset{a,b,\mathbf{w}}{\text{minimize}} \quad \mathrm{RMSE}(\mathbf{w},\, a,\, b) \\ &\text{subject to} \quad \sum_{j=1}^{J} w_j = 1 \quad \text{and} \quad w_j > 0, \quad j = 1, \ldots, J. \end{aligned} \tag{F.9}$$

This optimization problem is non-convex and we are not aware of a method to solve it analytically. Instead, we apply the MATLAB Optimization Toolbox to numerically find solutions which are locally optimal.

3 Experimental Data

We use two datasets of measured intelligibility to investigate the fitting of BIFs according to (F.9), and to compare the resulting prediction performance with that of the original STOI measure.

3.1 The "Kjm" dataset

The first dataset was used in the initial evaluation of the STOI measure [9] and is described in detail in [19]. For this dataset, intelligibility was measured for 15 normal hearing Danish subjects using the Dantale II corpus [20]. Measurements were carried out for 1) four noise types: Speech Shaped Noise (SSN), café noise, bottling factory noise and car noise 2) processing by two types of binary masks, Ideal Binary Masks (IBMs) and Target Binary Masks (TBMs), 3) eight different threshold values for binary mask generation and 4) three different SNRs. Since IBMs and TBMs are identical for SSN, there are only seven combinations of noise types and binary masks. The three SNRs were chosen individually for each noise type. Intelligibility was measured for a total of: 15 subjects $\times$ 7 noise/mask combinations $\times$ 8 RC values $\times$ 3 SNRs $\times$ 2 repetitions $\times$ 5 words/sentence $=$ 25200 words. By averaging performance across subjects, repetitions and words, we obtain measured intelligibility for 168 conditions. The authors of [19] have kindly supplied both clean and degraded audio files for the conditions.

For this study, the data is divided into eight subsets such as to investigate the BIFs arising from fitting to different types of data. Firstly, the dataset is divided into four subsets depending on noise type. Secondly, the dataset is divided according to the three SNR conditions (low, medium and high). Lastly, one subset is defined to include all the data. We refer to these subsets with the label "Kjm".

3.2 The "S&S" dataset

The second dataset [21] was collected in an effort to derive BIFs for the AI. Speech intelligibility was measured for 8 normal hearing subjects using a recording of the CID W-22 word lists. Measurements were carried out for 1) HP and LP filtered speech masked by SSN, 2) 21 filter cutoff frequencies and 3) 10 different SNRs. SNRs were uniformly spaced in 2 dB intervals between -10 and +8 dB. In total, this amounts to 2 filter types (HP/LP) $\times$ 21 cutoff frequencies $\times$ 10 SNRs $=$ 420 conditions. However, some conditions were skipped because intelligibility was almost zero, and therefore only 308 conditions were measured [21]. The results are shown in Fig. F.1.

It has not been possible to obtain either clean or degraded speech for the conditions of this experiment. Nor has it been possible to obtain recordings of

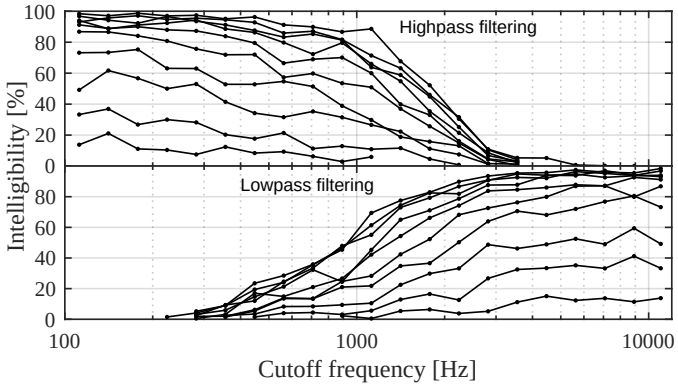


Fig. F.1: Replotted experimental results, as reported in tables 2–3 of [21]. The top plot shows measured intelligibility of HP filtered noisy speech versus cutoff frequency. Each line represents measurements at a particular SNR. The bottom plot shows the same type of results for LP filtering.

the CID W-22 word lists. We therefore recreated similar stimuli as accurately as possible, in order to allow for computing STOI scores. To this end, 150 random sentences were selected from the TIMIT database [22] and concatenated. Both HP and LP filtering was carried out using 512th order linear phase Finite Impulse Response (FIR) filters, designed using the windowing method. SSN was generated by filtering white noise such as to have the same long time spectrum as the TIMIT sentences. The concatenated, non-filtered, TIMIT sentences were used as a clean reference signal, ($x(t)$), while filtered speech, mixed with SSN, was used as degraded speech ($y(t)$). The SNR is defined to be the energy ratio of speech and noise *before* filtering the speech (as in [21]).

We define three divisions of this dataset: 1) the conditions with HP filtering, 2) the conditions with LP filtering, and 3) all the data. We refer to these subsets with the label "S&S". We also define one set of data, "Kjm+S&S", which includes all data from both experiments.

3.3 ANSI SII- and Uniform BIFs

In addition to BIFs fitted with (F.9), we include two additional BIFs: 1) The BIF specified for use with the SII in Table 3 of [2]. Linear interpolation was used to determine a BIF for the exact center frequencies of the one-third octave bands of the STOI measure. This BIF, shown in Fig. F.2, places increased weight on the higher frequency bands, as compared to the uniform BIF. 2) A uniform BIF, as used in the original STOI measure [9],

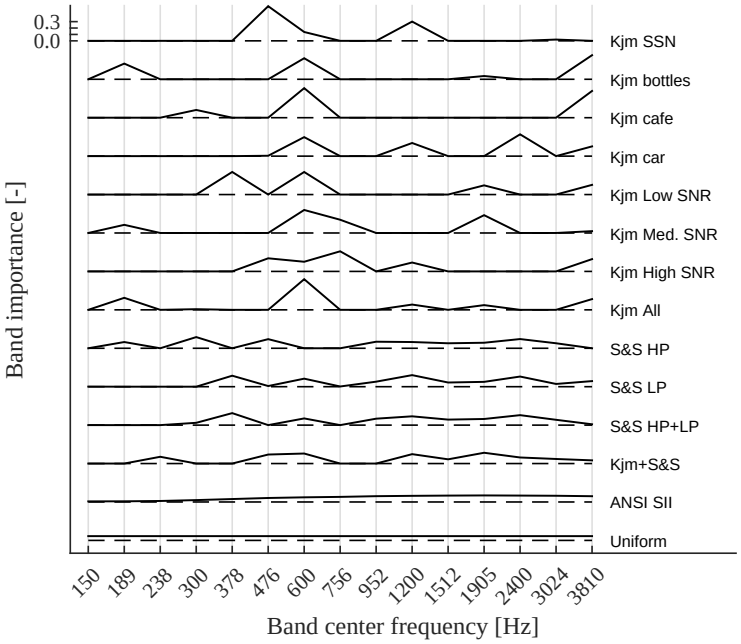


Fig. F.2: Fitted BIFs for eight subsets of the ”Kjm’ data, three subsets of the ”S&S” data, one set including data for both experiments, as well as two non-fitted standard BIFs. The scaling of the vertical axes is the same for all BIFs.

i.e. $w_j = 1/J, \quad j = 1, \ldots, J.$

4 Results and Discussion

BIFs were fitted to the defined subsets of data by finding local minima for (F.9), using the `fminsearch`-solver in the MATLAB Optimization Toolbox[1]. The resulting BIFs are shown in Fig. F.2. Most strikingly, all BIFs fitted to subsets of the ”Kjm”-data place the majority of the weight on few frequency bands. The heavily weighted bands are not the same across the BIFs (except for band 7, which is consistently weighted strongly by all ”Kjm”-BIFs except the one fitted to the SSN conditions). Such solutions could indicate some degree of overfitting, and it should be remarked that the smaller subsets of the ”Kjm”-data involve only 24, 48 or 56 data points, to which 17 parameters are fitted (i.e. a, b and $\mathbf{w} \in \mathbb{R}^{15\times 1}$). However, the full set of all 168 data points of the ”Kjm”-data results in a BIF with similar

[1]The default solver was initialized 100 times with random starting values, and the best solution across these was used.

properties. It should also be noted that while the BIFs place most weight on a few bands, these few bands are generally spread out across the entire frequency range. Another explanation of the sparse BIFs could therefore be that the values of $\tilde{d}_j$ are highly correlated for adjacent bands, and thus supply redundant information. It is possible that smoother BIFs can be obtained by adding some form of regularization to (F.9).

The BIFs fitted to the subsets of the "S&S"-data appear much smoother than those fitted to the "Kjm"-data. At the same time, the "S&S"-BIFs are similar to one-another. Especially the BIFs fitted to the "S&S LP"- and the "S&S LP+HP"-subsets show some similarity to the SII BIF, by weighting the higher frequency bands slightly higher than the lower ones. The joint set of data from both experiments, "Kjm+S&S", leads to a BIF which is quite similar to the one fitted to the "S&S HP+LP"-data. This could indicate that the RMSE of the "S&S"-data is more sensitive to differences in BIFs, and that this dataset therefore ends up having the most influence on the optimal BIF. This is not surprising, as the "S&S"-data is designed specifically with the purpose of containing as much information as possible about which frequency bands are important to speech intelligibility (i.e. to facilitate the derivation of BIFs).

We evaluate the performance of all 14 BIFs on all 12 subsets of data, using two different performance metrics: 1) RMSE, and 2) Kendall's tau. The results are shown in color-coded tables in Fig. F.3.

We first consider performance in terms of RMSE, as given by the top plot of Fig. F.3. Each fitted BIF is optimized to minimize the RMSE on one particular dataset. This is seen in Fig. F.3 as a diagonal with high performance, projecting from the lower left corner. It can be noted that BIFs fitted on one subset of the "Kjm"-data often leads to a low RMSE when used on another subset of the "Kjm"-data, with some exceptions. This contradicts the notion of overfitting being a major problem with the small subsets of the "Kjm"-data. A similar observation holds for the "S&S"-data, where rather good performance is obtained regardless of which BIF is evaluated for what subset of data. In general it appears that lower RMSE can be obtained on the "S&S"-data, which suggests that this dataset contains either less statistical variation or less varied combinations of noise and processing. When using BIFs fitted to the "Kjm"-data for predictions of the "S&S"-data, and vice versa, performance is mostly low. This suggests some fundamental difference between the two datasets, caused e.g. by differences in target speech material. However, the combined "Kjm+S&S"-BIF manages to obtain good performance across all subsets of both sets of data. The uniform- and SII BIFs also obtain decent performance across most conditions, especially when considering that these are not fitted to any of the available data. With the exception of the "Kjm+S&S"-BIF, the uniform BIF, as used in the original STOI measure, has the smallest maximum RMSE (i.e. the highest number of the row: 14.3%). However, RMSE measured on all the available data combined, as shown in

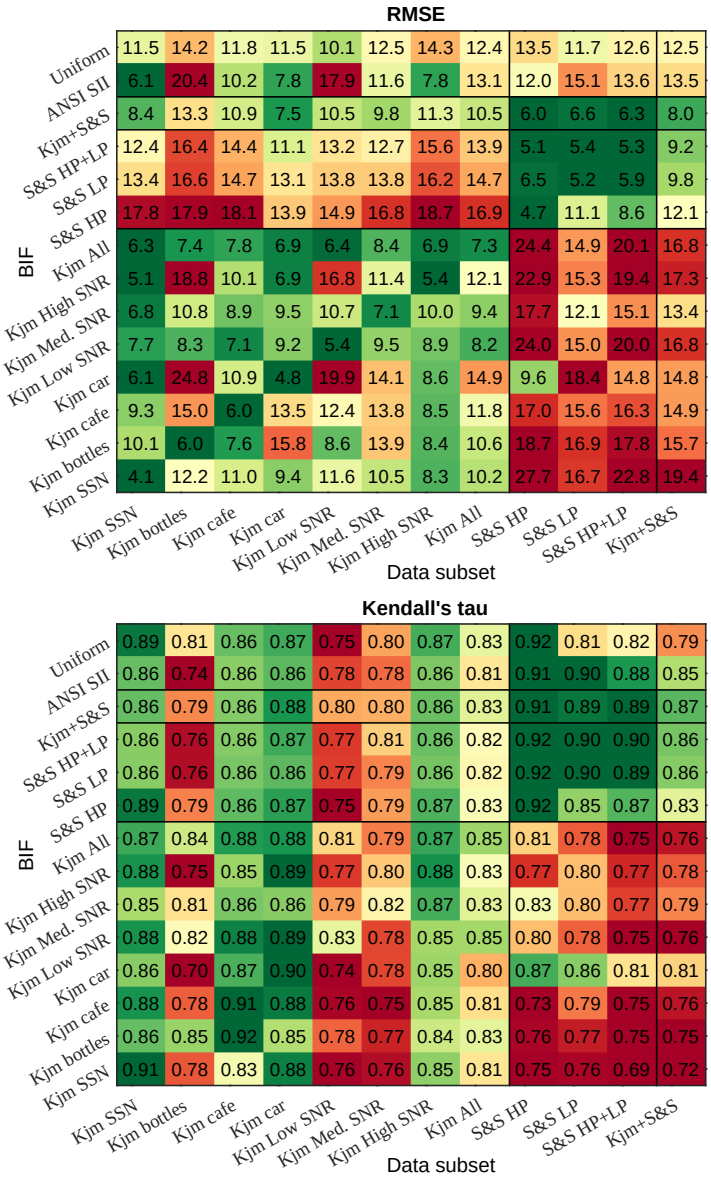


Fig. F.3: Cross evaluation of all the BIFs with the 12 defined subsets of data. Each row shows the performance for one BIF, when evaluated on the different subsets of data. Each column shows the performance of the different BIFs when evaluated for one particular data subset. The top plot shows RMSE in % and the bottom plot shows Kendall's tau. Red colors indicate poorer than average performance and green colors indicate better than average performance.

the rightmost column, is lowest for the "Kjm+S&S"-BIF, by a considerable margin. All BIFs fitted on the "Kjm"-data lead to quite poor performance when evaluated for the combined data, while the "S&S"-BIFs lead to much better performance. This should be viewed in light of the fact that the "S&S"-dataset is almost twice as big as the "Kjm"-dataset and therefore weighs more in the combined performance evaluation.

One can argue that it is unfair to fit BIFs to data from one listening experiment and validate it on data from another, because the speech material may have different degrees of complexity and the different groups of subjects may not perform equally well. These factors are, to a large extent, modeled by the parameters a and b, which control the mapping from STOI measure to predicted intelligibility in percent. The bottom plot in Fig. F.3 shows performance in terms of Kendall's tau. This statistic is interesting because it depends only on the extent to which predictions are correctly ordered, and is therefore independent of a and b. Here, we also see that fitting and testing with the same set of data gives improved performance, but to a somewhat smaller extent than what is the case with the RMSE which is directly optimized in (F.9). It is also seen that poor performance results when fitting BIFs on the "Kjm"-data and evaluating on the "S&S"-data, as was also the case when measuring performance in terms of RMSE. However, the opposite is not the case: fitting BIFs on the "S&S"-data and evaluating on the "Kjm"-data leads to performance which is almost as good as what is obtained when fitting with the "Kjm All"-set. This contrasts the results seen when evaluating with RMSE, and may indicate that a and b are important for fitting details about the specific experiment, and are not transferable from one experiment to another. On the other hand, this result also indicates that the BIF, $\mathbf{w}$, fitted on the "S&S"-data actually generalizes well to the "Kjm"-data. Overall, the BIF fitted to the "Kjm+S&S"-set performs better than the uniform BIF, in terms of Kendall's tau, when evaluated on the "Kjm+S&S"-set. However, this difference seems to stem mainly from the "S&S LP"-conditions. The other conditions do not indicate that performance is improved considerably above that of the uniform BIF.

5 Conclusions

We have investigated the use of Band Importance Functions (BIFs) in the Short-Time Objective Intelligibility (STOI) measure. BIFs were fitted to several different datasets of measured intelligibility, such as to minimize the Root-Mean-Square Error (RMSE). This can decrease prediction RMSE substantially in comparison with the uniform weighting of frequency bands normally used in the STOI measure. However, when cross evaluation was carried out between different sets of data, or when performance was measured us-

ing Kendall's tau, the use of BIFs appeared to result in neither a large or a consistent improvement in performance across the evaluated conditions. It is therefore not possible to say from this limited study, whether the improved average performance generalizes to other conditions. Across most of the evaluated conditions, it appears that the uniform BIF, as applied in the original STOI measure, is nearly optimal.

6 Acknowledgements

This work was funded by the Oticon Foundation and the Danish Innovation Foundation.

References

[1] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[2] A. S. S3.5-1997, *Methods for Calculation of the Speech Intelligibility Index*, ANSI Std. S3.5-1997, 1997.

[3] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.

[4] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, Dec. 2006.

[5] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, Apr. 2005.

[6] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 331–342, Apr. 2006.

[7] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension and evaluation of a binaural speech intelligibility model," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2479–2497, Dec. 2010.

[8] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 134, no. 1, pp. 436–446, Jul. 2013.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for inteligibility prediction of time-frequency weighted noisy speech," *IEEE Tran. on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[10] K. Smeds, A. Leijon, F. Wolters, A. Hammarstedt, S. Båsjö, and S. Hertzman, "Comparison of predictive measures of speech recognition after noise reduction processing," *J. Acoust. Soc. Am.*, vol. 136, no. 3, pp. 1363—-1374, Sep. 2014.

[11] S. Jørgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acta Acustica United with Acustica*, vol. 101, pp. 1016–1025, 2015.

[12] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[13] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Predicting the intelligibility of noisy and non-linearly processed binaural speech," *Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 1908–1920, 2016.

[14] L. Lightburn and M. Brookes, "A weighted STOI intelligibility metric based on mutual information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China: IEEE, Mar. 2016, pp. 5365–5369.

[15] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[16] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, US: IEEE, Mar. 2017, pp. 5085–5089.

[17] J. M. Kates, "Improved estimation of frequency importance functions," *J. Acoust. Soc. Am.*, vol. 134, no. 5, pp. EL459–EL464, Nov. 2013.

[18] J. Ma, Y. H. Philipos, and C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3387–3405, May 2009.

[19] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.

[20] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *International Journal of Audiology*, vol. 42, no. 1, pp. 10–17, Jan. 2003.

[21] G. A. Studebaker and R. L. Sherbecoe, "Frequency-importance and transfer functions for recorded CID W-22 word lists," *Journal of Speech and Hearing Research*, vol. 34, pp. 427–438, Apr. 1991.

[22] DARPA, "TIMIT, acoustic-phonetic continuous speech corpus."

Paper G

Refinement and Validation of the Binaural Short Time Objective Intelligibility Measure for Spatially Diverse Conditions

Asger Heidemann Andersen
Jan Mark de Haan
Zheng-Hua Tan
Jesper Jensen

Submitted to
Speech Communication.

© 2017 Elsevier
The layout has been revised.

Abstract

Speech intelligibility prediction methods have recently gained popularity in the speech processing community as supplements to time consuming and costly listening experiments. Such methods can be used to objectively quantify and compare the advantage of different speech enhancement algorithms, in a way that correlates well with actual speech intelligibility. One such method is the short-time objective intelligibility (STOI) measure. In a recent publication, we proposed a binaural version of the STOI measure, based on a modified version of the equalization cancellation (EC) model. This measure was shown to retain many of the advantageous properties of the STOI measure, while at the same time being able to predict intelligibility correctly in conditions involving both binaural advantage and non-linear signal processing. The biggest prediction errors were found for conditions involving multiple spatially distributed interferers. In this paper, we report results for a new listening experiment including different mixtures of isotropic and point source noise. This exposes that the binaural STOI has a tendency to overestimate the intelligibility in conditions with spatially distributed interferes at low signal to noise ratios (SNRs). This condition-dependent error can make it difficult to compare intelligibility across different acoustical conditions. We investigate the cause of this upward bias, and propose a correction which alleviates the problem. The modified method is evaluated with five datasets of measured intelligibility, spanning a wide range of realistic acoustic conditions. Within the tested conditions, the modified method yields very accurate predictions, and entirely alleviates the aforementioned tendency to overestimate intelligibility in conditions with spatially distributed interferers.

1 Introduction

Speech Intelligibility Prediction (SIP) algorithms aim to predict the intelligibility of noisy or degraded recordings of speech. Such algorithms were first studied early in the twentieth century as a means to quantify the intelligibility of speech transmitted via telephone [1, 2]. Recently, SIP has become an increasingly popular tool for objectively assessing intelligibility within applications such as speech enhancement [3–8], architectural acoustics [9, 10], and telecommunications [11].

SIP algorithms can be separated in two groups: intrusive ones and non-intrusive ones. Intrusive SIP algorithms predict intelligibility using some representation of the degraded signal as well as a clean reference signal, while non-intrusive SIP algorithms require only the degraded signal. The focus of this paper is on intrusive SIP algorithms.

The first widely known SIP algorithm was the Articulation Index (AI), which assumes speech intelligibility to be proportional to a weighted sum of long-term Signal to Noise Ratios (SNRs) across different frequency bands [2].

A theoretical justification of the AI is provided in [12], which shows it to be an estimate of the channel capacity from Shannon's information theory. A refined and standardized version of the AI is known as the Speech Intelligibility Index (SII) [13]. While the SII has proven successful in many respects, conditions exist, where it is unable to provide accurate predictions. Therefore, many alternative SIP algorithms have been proposed, often inspired by the SII, but aiming to extend the domain of signals for which accurate predictions can be made (some overviews are provided by [4, 9, 11, 14–17]). Here, we provide a brief overview of SIP algorithms which extend the domain of the SII in three different ways: by taking into account fluctuating interferers, non-linear processing and binaural advantage, respectively:

- **Fluctuating interferers:** Speech intelligibility may be substantially higher in the presence of modulated or interrupted noise, than in the presence of stationary noise at a similar level [18]. This suggests that humans are capable of effectively collecting information from short instants with high SNR. The SII relies on long-time signal statistics, and is therefore unable to take such an effect into account [19]. This ability is, however, important to consider, as many real-world noise signals are highly non-stationary (e.g. competing talkers). One solution to this problem is given by the Extended SII (ESII) [19, 20], which computes the SII in short time windows, and averages contributions across time. The ESII has shown high prediction accuracy in conditions with interrupted and sinusoidally modulated Speech Shaped Noise (SSN) at various frequencies [20]. The principle of averaging contributions across short time segments has repeatedly been used for the same purpose in other predictors [3, 14, 21]. An alternative model, proposed in [22], predicts intelligibility from the fraction of the time-frequency plane for which the SNR exceeds a fixed threshold.

- **Non-linear processing:** Since the SII bases predictions on SNRs, it is not applicable in conditions, where the noisy speech signal has been non-linearly processed or distorted, because the SNR is not well defined in this case. Since SIP is often used for evaluating the performance of non-linear speech enhancement algorithms, it is desirable to have SIP algorithms which can handle such processing. One such approach is the Coherence SII (CSII), which computes the SII based on coherence rather than SNR [23]. This measure has shown good prediction performance for distorted speech [23] and noise reduction processing [17]. Another approach is the frequency-weighted segmental SNR (fwSNRseg) [14, 15], which uses the difference between clean and degraded signal sub-band envelopes as an estimate of the noise envelope, and further uses this to compute an SII-like measure. A somewhat different model is the speech-based EPSM (sEPSM) [24], which is based

on a modulation domain model of human hearing [25]. The sEPSM performs well at predicting the impact of spectral subtraction algorithms [24] and noisy speech transmitted via telephone [11]. It has also been proposed in a short-time version which can handle fluctuating interferers [21]. Recently, the Short-Time Objective Intelligibility (STOI) measure has gained considerable popularity, due to its simplicity and high performance in a range of key scenarios such as different noise types [3], noise reduction processing [3, 17], hearing aid processing [4] and noisy speech transmitted via telephone [11]. The Extended STOI (ESTOI) measure [26] has later been proposed with the aim og improving prediction performance in conditions with fluctuating interferers.

- **Binaural advantage:** The SII considers speech presented to one ear only or, similarly, identical speech signals presented to both ears. In real-world situations, a large binaural advantage may be obtained, because the acoustics of the head causes different sound signals to reach the left and right ears. Binaural advantage can have a large impact on speech intelligibility [27], and is increasingly considered in speech enhancement applications [28, 29]. It is therefore a highly desirable property of a SIP algorithm to predict this effect. The binaural advantage can be seen as originating from two separate sources: 1) the human head casts an acoustical shadow, that leads to Interaural Level Differences (ILDs), dependent on the source location with respect to the ears [30]. In some acoustical conditions, this causes the SNR to be higher at one ear than at the other. The potential ability of the brain to listen to the ear with the higher SNR may lead to a considerable advantage, which we refer to as the ”better-ear” advantage. 2) The distance between the ears causes Interaural Time Differences (ITDs) which depend on the location of the source [30]. Experiments have indicated that the brain can exploit ITDs to effectively segregate sound signals from different spatial locations [27, 31]. We refer to this as an advantage due to ”binaural unmasking”. The Equalization Cancellation (EC) stage [32, 33] models binaural advantage, by assuming that the brain uses delays and gain adjustments to align any interferer components in the left and right ears, and thereafter subtracts the two signals from one-another to cancel the interferers. The Binaural Speech Intelligibility Measure (BSIM) [34, 35] combines the SII and the EC stage in order to predict intelligibility including binaural advantage. This is done by optimizing the delay and gain parameters in the EC stage to maximize the SII. A short-time version, the short-time BSIM (stBSIM), is proposed in [35]. An approach similar to that of the BSIM is followed in [36], with a short-time version following in [37]. A different approach comes from [38–40], which

proposes to compute the SII from the better ear, and adding binaural advantage as estimated by an expression from [41]. A binaural version of the STOI measure is proposed in [42] and further improved in [43, 44]. The improved version is referred to as the Deterministic BSTOI (DBSTOI) measure. This is based on extending the STOI measure with an EC stage, similar to how the BSIM extends the SII. The DBSTOI measure retains many of the favorable properties of the STOI measure while being able to predict binaural advantage [44]. A binaural version of the multi-resolution sEPSM (mr-sEPSM), the Binaural sEPSM (B-sEPSM), is proposed in [45]. This too, is based on an EC stage. Both the DBSTOI measure and the B-sEPSM are short-time methods, and both can handle non-linearly processed speech.

The evaluation of the DBSTOI measure in [44] generally shows high prediction accuracy in the tested conditions. The largest prediction errors are seen for conditions with spatially distributed interferers such as isotropic noise or multiple point-like interferers in different spatial locations. It is important that a binaural SIP algorithm works well in conditions with spatially distributed interferers which are common in practice (e.g. at a cocktail party or in a reverberant room). Furthermore, many speech enhancement algorithms modify the binaural cues of the processed signals, and thereby modify the conveyed acoustical scene. An extreme example of this is binaural beamforming which may transform the binaural signal into a diotic one, thus collapsing all sound sources towards the front of the listener [28, 29]. It is important that a binaural SIP algorithm is able to correctly predict the impact of such processing.

In this paper we carry out a thorough investigation of the DBSTOI measure for distributed interferers. In Section 2 we introduce five datasets of measured intelligibility, three of which include different combinations of spatially distributed interferers. We introduce the datasets at this early point, as we refer to them throughout the paper. In Section 3 we 1) provide an outline of the DBSTOI measure, 2) investigate the problems arising with the DBSTOI measure in conditions with spatially distributed interferers as well as the cause of these, and 3) propose a modified measure, the Modified BSTOI (MBSTOI) measure, which solves the observed problems. We also propose a computationally simpler better-ear version, the better-ear MBSTOI (beMBSTOI), which does not model binaural unmasking. We do so, mainly to investigate the relative contributions of better-ear advantage and binaural unmasking to predicted intelligibility. In Section 4, we compare the DBSTOI, MBSTOI, and beMBSTOI measures, as well as the B-sEPSM, using the five previously mentioned datasets. Section 5 concludes upon our findings.

Table G.1: Summary of conditions. Abbreviations are as follows: SSN (speech shaped noise), BFHN (bottling factory hall noise), ISTS (international speech test signal), ISO (cylindrically isotropic SSN), IBM (ideal binary mask processing), un. (unprocessed), bil. (bilateral), bin. (binaural), MVDR (minimum variance distortionless response beamforming), MWF (multichanncel Wiener filtering), Kukl. (algorithms proposed in [46]), Braun (algorithms proposed in [47]).

Cond.	Interferer	Interferer location(s)	Proc.
D_1-1	SSN	$-160°$	–
D_1-2	SSN	$-115°$	–
D_1-3	SSN	$-80°$	–
D_1-4	SSN	$-40°$	–
D_1-5	SSN	$20°$	–
D_1-6	SSN	$0°$	–
D_1-7	SSN	$40°$	–
D_1-8	SSN	$80°$	–
D_1-9	SSN	$140°$	–
D_1-10	SSN	$180°$	–
D_2-1	SSN	$-115°$	IBM
D_2-2	SSN	$0°$	IBM
D_2-3	SSN	$20°$	IBM
D_2-4	BFHN	$-115°$	–
D_2-5	BFHN	$0°$	–
D_2-6	BFHN	$20°$	–
D_2-7	BFHN	$-115°$	IBM
D_2-8	BFHN	$0°$	IBM
D_2-9	BFHN	$20°$	IBM
D_3-1	SSN	ISO	–
D_3-2	SSN	$\{\pm 115°, 180°\}$	–
D_3-3	ISTS	$\{\pm 115°, 180°\}$	–
D_3-4	SSN	$\{30°, 180°\}$	–
D_3-5	ISTS	$\{30°, 180°\}$	–
D_3-6	SSN	ISO	Beamforming
D_3-7	SSN	$\{\pm 115°, 180°\}$	Beamforming
D_3-8	ISTS	$\{\pm 115°, 180°\}$	Beamforming
D_3-9	SSN	$\{30°, 180°\}$	Beamforming
D_3-10	ISTS	$\{30°, 180°\}$	Beamforming
D_4-1	ISTS	$\{\pm 90°, \pm 135°, 180°\}$	Un. front mics
D_4-2	ISTS	$\{\pm 90°, \pm 135°, 180°\}$	Un. left front mic
D_4-3	ISTS	$\{\pm 90°, \pm 135°, 180°\}$	Bil. MVDR Kukl.
D_4-4	ISTS	$\{\pm 90°, \pm 135°, 180°\}$	Bil. MWF Kukl.
D_4-5	ISTS	$\{\pm 90°, \pm 135°, 180°\}$	Bin. MVDR Braun
D_4-6	ISTS	$\{\pm 90°, \pm 135°, 180°\}$	Bin. MWF Braun
D_4-7	ISTS	$\{\pm 90°, \pm 135°, 180°\}$	Bin. MVDR Kukl.
D_4-8	ISTS	$\{\pm 90°, \pm 135°, 180°\}$	Bin. MWF Kukl.
D_5-1	SSN	$0°$	–
D_5-2	SSN	$\frac{2}{3}\{0°\} + \frac{1}{3}\{\text{ISO}\}$	–
D_5-3	SSN	$\frac{1}{3}\{0°\} + \frac{2}{3}\{\text{ISO}\}$	–
D_5-4	SSN	ISO	–
D_5-5	SSN	$\frac{2}{3}\{\text{ISO}\} + \frac{1}{3}\{-115°\}$	–
D_5-6	SSN	$\frac{1}{3}\{\text{ISO}\} + \frac{2}{3}\{-115°\}$	–
D_5-7	SSN	$-115°$	–

2 Experimental Data

In this section we describe five datasets of measured intelligibility, used throughout the paper for evaluating SIP algorithms. We use existing datasets from [42], [44], and [46], as well as a new dataset, that has been collected for evaluating the DBSTOI measure in isotropic noise. In combination, these datasets span a wide range of realistic conditions. We denote the datasets as D_1, D_2, D_3, D_4 and D_5, respectively. The acoustical conditions of the five datasets are listed in Table G.1.

2.1 D_1: Point Noise Sources

This dataset was initially described in [42], and features conditions with a frontal interferer, masked by a single point-like source of SSN located at different azimuths in the horizontal plane. Such conditions are acoustically simple but serve well to demonstrate binaural advantage. The experiment included 10 Danish test subjects who reported normal hearing. They were presented with noisy speech stimuli via headphones in quiet. An anechoic listening environment was simulated using the CIPIC Head Related Transfer Functions (HRTFs) [48]. Sentences from the Dantale II corpus [49] were presented from the front, masked by an SSN interferer placed at some az-

imuth in the horizontal plane. Subjects were presented with one sentence at a time and asked to repeat it as accurately as possible. The number of correctly repeated words was noted by the experimenter. Speech intelligibility was measured for ten different interferer locations (see Table G.1), each at six SNRs, uniformly spaced by 3 dB and centered around a rough estimate of the Speech Reception Threshold (SRT)[1] (manually obtained by the experimenter, by listening to samples of the stimuli). Three sentences were scored for each subject/location/SNR, resulting in the scoring of (10 subjects) $\times$ (10 interferer locations) $\times$ (6 SNRs) $\times$ (3 repetitions) $=$ (1800 sentences).

2.2 D_2: Point Sources and Binary Mask Processing

This dataset was initially described in [43], and contains conditions similar to dataset D_1, but adds Ideal Binary Mask (IBM) processing [50] and realistic environmental noise. This dataset is included in order to investigate whether SIP algorithms can predict the combined impact of binaural advantage and non-linear processing. Measurements were carried out for 14 subjects who reported normal hearing. Three representative interferer azimuths were chosen from the conditions of dataset D_1 (0°, 20°, and -115°). Measurements were made for three different variations of each of these three conditions:

- **Conditions 1–3:** use an SSN interferer as in dataset D_1. However, the stimuli were IBM processed before being presented to the subjects: the speech and noise signals were decomposed with a short-time Discrete Fourier Transformation (DFT) and noisy DFT-units with an SNR of less than 0 dB were attenuated by 10 dB [44]. This processing was carried out independently on the left and right ear signals, e.g. to simulate a hearing aid with a powerful noise reduction system.
- **Conditions 4–6:** use an environmental noise recording in place of the SSN interferer: Bottling Factory Hall Noise (BFHN) [51]. BFHN is a recording of heavy machinery and bottles rattling on a conveyor belt. No IBM processing was applied.
- **Conditions 7–9:** are identical to conditions 4–6, but use IBM processing as in conditions 1–3.

Sentence material and scoring was carried out as for D_1, except for the fact that subjects did not verbally repeat the sentences they heard, but rather entered them on a computer screen by selecting each word from a list of 10

[1] We define the SRT as the SNR at which the intelligibility is 50%. The SNR is computed as the ratio of average target power to average interferer power, both measured at the respective sources.

possibilities[2]. This resulted in the scoring of (14 subjects) $\times$ (9 conditions) $\times$ (6 SNRs) $\times$ (3 repetitions) $=$ (2268 sentences).

2.3 D_3: Multiple Interferers and Beamforming

This dataset was initially described in [42], and includes different combinations of multiple interferers, isotropic noise and beamforming. It is therefore well suited for evaluating, how SIP algorithms cope with changing acoustical conditions and processing. Intelligibility was measured for 10 subjects, who reported normal hearing, using the same sentence material and collection procedure as dataset D_1. In all the 10 conditions, speech was presented from the front. Otherwise, the conditions are as follows:

- **Condition 1:** speech is presented together with cylindrically isotropic SSN.
- **Condition 2:** SSN interferers are placed in 115°, 180° and -115° azimuth.
- **Condition 3:** is acoustically identical to condition 2, but uses the International Speech Test Signal (ISTS) instead of SSN. ISTS is a speech-like, but unintelligible, signal, obtained by concatenating short segments of speech in different languages [53].
- **Condition 4:** SSN interferers are placed in, 30° and -180° azimuth.
- **Condition 5:** is acoustically identical to condition 4, but uses the ISTS instead of SSN.
- **Conditions 6–10:** are identical to conditions 1–5, but present subjects with the output of a time-invariant, two-microphone Minimum Variance Distortionless Response (MVDR) beamformer. This was accomplished by using HRTFs measured for the two microphones of a behind-the-ear hearing aid.

This resulted in the scoring of (10 subjects) $\times$ (10 conditions) $\times$ (6 SNRs) $\times$ (3 repetitions) $=$ (1800 sentences).

2.4 D_4: Reverberation and Beamforming

This dataset was initially described in [46], and consists of conditions with multiple spatially separated interferers and different types of beamforming. Some types of beamforming alter the binaural cues of the input signals,

[2]The collection procedure is described in more detail in [44]. A comparison of the two approaches for measuring intelligibility is conducted in [52], showing them to yield highly similar results.

and this dataset allows for evaluating how SIP algorithms cope with such changes. Measurements were made for 20 subjects, using the same speech material and procedure as in dataset D_2. Speech was presented from the front with reverberation simulated by use of a Binaural Room Impulse Response (BRIR) measured in a large cellar. Five ISTS interferers were placed at azimuths of $\pm 90^\circ$, $\pm 135^\circ$ and 180°, respectively, using BRIRs measured in those locations in the same cellar. The interferers were fixed at the same power as the target speaker. Intelligibility was modified by synthetically altering the Direct to Reverberant Ratio (DRR) of the target speaker. Intelligibility was measured for four different DRRs and for eight different combinations of hearing aid beamforming and noise reduction. Details on the applied processing are given in [46]. In total, intelligibility was scored for (20 subjects) $\times$ (8 conditions) $\times$ (4 DRRs) $\times$ (5 repetitions) $=$ (3200 sentences).

2.5 D_5: Isotropic and Point Source Interferers

To gain even more knowledge about the performance of the DBSTOI measure in environments with distributed interferers, we collected an additional dataset with different combinations of cylindrically isotropic and point source SSN interferers. Specifically, we measured intelligibility for frontal speech, using the same speech material and procedure as that of dataset D_2. The speech was masked by interferer signals of the following form:

$$\mathbf{n}(t;\,\alpha) = \begin{cases} \sqrt{1-\alpha}\,\mathbf{n_0}(t) + \sqrt{\alpha}\,\mathbf{n}_{\text{iso}}(t), & \text{if } 0 < \alpha < 1. \\ \sqrt{2-\alpha}\,\mathbf{n}_{\text{iso}}(t) + \sqrt{\alpha-1}\,\mathbf{n}_{\mathbf{-115}}(t), & \text{if } 1 < \alpha < 2, \end{cases} \tag{G.1}$$

where $\mathbf{n_0}(t) : \mathbb{R} \to \mathbb{R}^{2\times 1}$ is the binaural signal arising from a point source of SSN directly in front of the listener, while $\mathbf{n}_{\text{iso}}(t)$ is the binaural signal arising from cylindrically isotropic SSN, and $\mathbf{n}_{\mathbf{-115}}(t)$ is the binaural signal arising from a point source of SSN at -115° from the front. Intelligibility was measured for seven linearly spaced values of α: 0, $1/3$, $2/3$, 1, $4/3$, $5/3$ and 2, each for eight SNRs linearly spaced by 3 dB, centered on a rough estimate of the SRT. Measurements were carried out with eight subjects, who reported normal hearing, using the same procedure as that described for dataset D_2. This resulted in the scoring of (8 subjects) $\times$ (7 conditions) $\times$ (8 SNRs) $\times$ (3 repetitions) $=$ (1344 sentences).

3 Methods

The purpose of this section is threefold: 1) to briefly outline the DBSTOI measure, which was initially proposed in [43, 44], 2) to highlight important disparities between measured intelligibility and predictions of the DBSTOI measure which arise in conditions with spatially distributed noise sources,

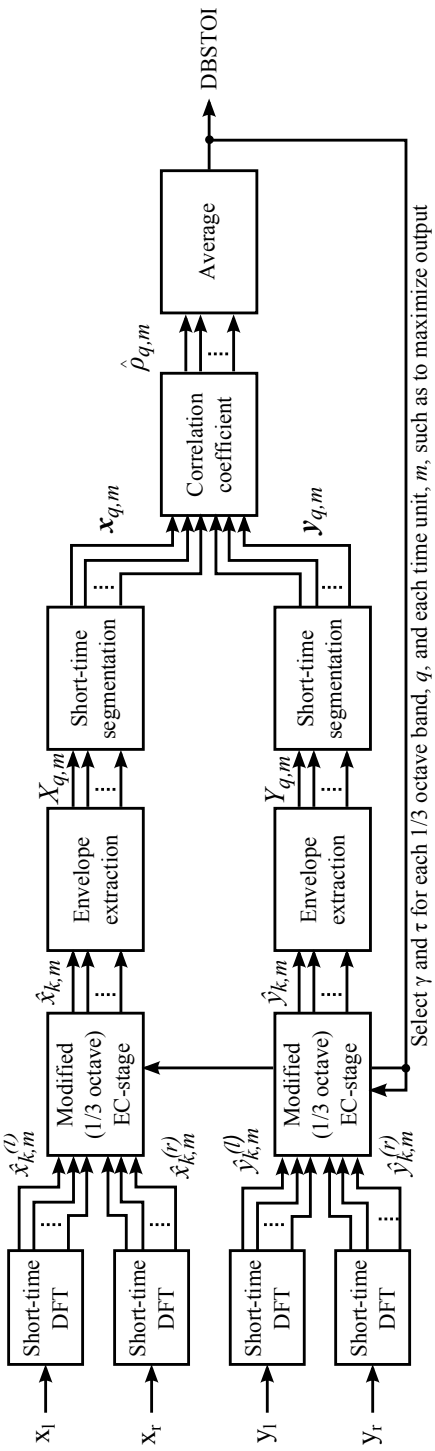


Fig. G.1: A block diagram of the DBSTOI measure. Figure taken from [44].

and 3) to propose a modification of the DBSTOI measure, which greatly diminishes these disparities.

3.1 The DBSTOI measure

The DBSTOI measure, introduced in [44], extends the STOI measure to enable prediction of binaural advantage. At the same time, it was shown that the DBSTOI measure mostly retains the favorable monaural prediction performance of the STOI measure [44]. An overview of the DBSTOI measure is given in Fig. G.1. To facilitate the following discussion, we give a compressed overview of the DBSTOI measure. A more thorough treatment can be found in [44].

As input signals, the DBSTOI measure assumes access to degraded signals, $y_l(t)$ and $y_r(t)$, recorded at the left and right ears of the listener, as well as the corresponding clean reference signals $x_l(t)$ and $x_r(t)$. These are analyzed with a short-time DFT, yielding coefficients, $\hat{y}_{k,m}^{(l)}$, $\hat{y}_{k,m}^{(r)}$, $\hat{x}_{k,m}^{(l)}$ and $\hat{x}_{k,m}^{(r)}$, where k is the frequency bin index and m is the frame index. The DFT is computed with the same parameters as in [3].

The DFT coefficients from the left and the right ear are combined using an EC stage [32, 33, 54], which models the binaural advantage obtained by having two ears, in situations with spatial separation between target and interferer [44]:

$$\hat{x}_{k,m} = \lambda_{k,m}\, \hat{x}_{k,m}^{(l)} - \lambda_{k,m}^{-1} \hat{x}_{k,m}^{(r)}, \tag{G.2}$$

$$\hat{y}_{k,m} = \lambda_{k,m}\, \hat{y}_{k,m}^{(l)} - \lambda_{k,m}^{-1} \hat{y}_{k,m}^{(r)}, \tag{G.3}$$

where

$$\lambda_{k,m} = 10^{(\gamma_{k,m}+\Delta\gamma_{k,m})/40} e^{j\omega(\tau_{k,m}+\Delta\tau_{k,m})/2}. \tag{G.4}$$

Here, $\gamma_{k,m}$ is a relative amplitude change between the left and right ear signals, and $\tau_{k,m}$ is a relative time shift. These are used to compute the complex scalar $\lambda_{k,m}$, which represents both time and amplitude shift. Both are introduced before subtracting the right ear coefficient from the left ear coefficient to obtain the EC stage output. The parameters $\gamma_{k,m}$ and $\tau_{k,m}$ are determined in order to maximize predicted intelligibility of the output. Informally, the purpose of $\gamma_{k,m}$ and $\tau_{k,m}$ can be interpreted as that of aligning any interferer or distortion components in the left and right ear signals, such that these are canceled when the signals are subtracted from one another [32]. In signal processing terms, the EC stage can be viewed as a beamformer. Independent noise sources, $\Delta\gamma_{k,m}$ and $\Delta\tau_{k,m}$, referred to as "jitter" [32], are added to $\gamma_{k,m}$ and $\tau_{k,m}$, to limit performance to be consistent with human performance. The

jitters are zero mean Gaussian with variances [54]:

$$\sigma_{\Delta\gamma}(\gamma_{k,m}) = \sqrt{2} \cdot 1.5 \text{ dB} \cdot \left(1 + \left(\frac{|\gamma_{k,m}|}{13 \text{ dB}}\right)^{1.6}\right) \quad [\text{dB}], \tag{G.5}$$

$$\sigma_{\Delta\tau}(\tau_{k,m}) = \sqrt{2} \cdot 65 \ \mu\text{s} \cdot \left(1 + \frac{|\tau_{k,m}|}{1.6 \text{ ms}}\right) \quad [\text{s}]. \tag{G.6}$$

The combined DFT coefficients are used to compute power envelopes, $X_{q,m}$ and $Y_{q,m}$, in $Q = 15$ one-third octave bands [44]:

$$\begin{aligned} X_{q,m} &= \sum_{k=k_1(q)}^{k_2(q)} \left|\hat{x}_{k,m}\right|^2 = \sum_{k=k_1(q)}^{k_2(q)} \left|\lambda\, \hat{x}_{k,m}^{(l)} - \lambda^{-1} \hat{x}_{k,m}^{(r)}\right|^2 \\ &= 10^{\frac{\gamma+\Delta\gamma}{20}} \sum_{k=k_1(q)}^{k_2(q)} \left|\hat{x}_{k,m}^{(l)}\right|^2 + 10^{-\frac{\gamma+\Delta\gamma}{20}} \sum_{k=k_1(q)}^{k_2(q)} \left|\hat{x}_{k,m}^{(r)}\right|^2 \\ &\quad - 2\text{Re}\left[\sum_{k=k_1(q)}^{k_2(q)} \hat{x}_{k,m}^{(l)*} \hat{x}_{k,m}^{(r)} e^{-j\omega_k(\tau+\Delta\tau)}\right] \\ &\approx 10^{\frac{\gamma+\Delta\gamma}{20}} X_{q,m}^{(l)} + 10^{-\frac{\gamma+\Delta\gamma}{20}} X_{q,m}^{(r)} \\ &\quad - 2\,\text{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)} X_{q,m}^{(c)}\right], \end{aligned} \tag{G.7}$$

where:

$$\begin{aligned} X_{q,m}^{(l)} &= \sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m}^{(l)}|^2, \\ X_{q,m}^{(r)} &= \sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m}^{(r)}|^2, \\ X_{q,m}^{(c)} &= \sum_{k=k_1(q)}^{k_2(q)} \hat{x}_{k,m}^{(l)*} \hat{x}_{k,m}^{(r)}, \end{aligned} \tag{G.8}$$

and where ω_k is the angular frequency of the k'th frequency bin, ω_q is the center angular frequency of the q'th third octave band, and $k_1(q)$ and $k_2(q)$ are, respectively, the lower and upper limits of the q'th one-third octave band. Similar power envelope samples, $Y_{q,m}$, are defined for the degraded signal. These envelopes are arranged in zero-mean vectors of $N = 30$ samples:

$$\mathbf{x}_{q,m} = [X_{q,m-N+1}, \ldots, X_{q,m}]^\intercal - \mathbf{1} \sum_{m'=m-N+1}^{m} \frac{X_{q,m'}}{N}, \tag{G.9}$$

where $\mathbf{1} \in \mathbb{R}^{N\times 1}$ is a vector of ones. Similar vectors, $\mathbf{y}_{q,m}$, are defined for the degraded signal, as well as for the other power envelope signals defined by (G.8): $\mathbf{x}^{(l)}_{q,m}$, $\mathbf{x}^{(r)}_{q,m}$, $\mathbf{x}^{(c)}_{q,m}$, $\mathbf{y}^{(l)}_{q,m}$, $\mathbf{y}^{(r)}_{q,m}$ and $\mathbf{y}^{(c)}_{q,m}$. Using (G.7), we then have [44]:

$$\begin{aligned}\mathbf{x}_{q,m} \approx\ & 10^{\frac{\gamma+\Delta\gamma}{20}}\mathbf{x}^{(l)}_{q,m} + 10^{-\frac{\gamma+\Delta\gamma}{20}}\mathbf{x}^{(r)}_{q,m} \\ & - 2\,\mathrm{Re}\left[e^{-j\omega_q(\tau+\Delta\tau)}\mathbf{x}^{(c)}_{q,m}\right].\end{aligned} \tag{G.10}$$

The STOI measure uses the average correlation between clean and degraded envelopes to predict intelligibility [3]. Assuming that $X_{q,m'}$ and $Y_{q,m'}$, for $m' = m-N+1, \ldots, m$, are samples of a wide sense stationary process, this can be interpreted as an estimate of the linear correlation coefficient:

$$\rho_{q,m} = \frac{E\left[(X_{q,m} - E\left[X_{q,m}\right])(Y_{q,m} - E\left[Y_{q,m}\right])\right]}{\sqrt{E\left[(X_{q,m} - E\left[X_{q,m}\right])^2\right] E\left[(Y_{q,m} - E\left[Y_{q,m}\right])^2\right]}}, \tag{G.11}$$

where the expectation is computed with respect to both the input signals, and the jitter in the EC stage. By estimating the correlation in the same manner as in the STOI measure, we arrive at [44]:

$$\bar{\rho}_{q,m} = \frac{E_\Delta\left[\mathbf{x}^{\mathsf{T}}_{q,m}\mathbf{y}_{q,m}\right]}{\sqrt{E_\Delta\left[\mathbf{x}^{\mathsf{T}}_{q,m}\mathbf{x}_{q,m}\right] E_\Delta\left[\mathbf{y}^{\mathsf{T}}_{q,m}\mathbf{y}_{q,m}\right]}}, \tag{G.12}$$

where $E_\Delta[\cdot]$ is the expectation with respect to the jitter. As shown in [44], the numerator of this expression may be approximated by:

$$\begin{aligned}&E_\Delta\left[\mathbf{x}^{\mathsf{T}}_{q,m}\mathbf{y}_{q,m}\right] \approx \\ &\quad(e^{2\beta}\mathbf{x}^{(l)\mathsf{T}}_{q,m}\mathbf{y}^{(l)}_{q,m} + e^{-2\beta}\mathbf{x}^{(r)\mathsf{T}}_{q,m}\mathbf{y}^{(r)}_{q,m})e^{2\sigma^2_{\Delta\beta}} \\ &+ \mathbf{x}^{(r)\mathsf{T}}_{q,m}\mathbf{y}^{(l)}_{q,m} + \mathbf{x}^{(l)\mathsf{T}}_{q,m}\mathbf{y}^{(r)}_{q,m} - 2e^{\sigma^2_{\Delta\beta}/2}e^{-\omega^2\sigma^2_{\Delta\tau}/2}\times \\ &\quad\left\{\left(e^{\beta}\mathbf{x}^{(l)\mathsf{T}}_{q,m} + e^{-\beta}\mathbf{x}^{(r)\mathsf{T}}_{q,m}\right) Re\left[\mathbf{y}^{(c)}_{q,m}e^{-j\omega\tau}\right]\right. \\ &+\ Re\left[e^{-j\omega\tau}\mathbf{x}^{(c)}_{q,m}\right]^{\mathsf{T}}\left.\left(e^{\beta}\mathbf{y}^{(l)}_{q,m} + e^{-\beta}\mathbf{y}^{(r)}_{q,m}\right)\right\} \\ &+ 2\left(Re\left[\mathbf{x}^{(c)H}_{q,m}\mathbf{y}^{(c)}_{q,m}\right] + e^{-2\omega^2\sigma^2_{\Delta\tau}} Re\left[\mathbf{x}^{(c)\mathsf{T}}_{q,m}\mathbf{y}^{(c)}_{q,m}e^{-j2\omega\tau}\right]\right),\end{aligned} \tag{G.13}$$

where:

$$\begin{aligned}\beta &= \frac{\ln(10)}{20}\gamma, \\ \sigma^2_{\Delta\beta} &= \left(\frac{\ln(10)}{20}\right)^2\sigma^2_{\Delta\gamma}.\end{aligned} \tag{G.14}$$

The values of $E_\Delta\left[\mathbf{x}_{q,m}^\intercal \mathbf{x}_{q,m}\right]$ and $E_\Delta\left[\mathbf{y}_{q,m}^\intercal \mathbf{y}_{q,m}\right]$ can be obtained from similar expressions (by exchanging all occurrences of $\mathbf{y}$ with $\mathbf{x}$, or vice versa, in (G.13)) [44]. The final DBSTOI measure is obtained by averaging the intermediate correlation coefficients across time and one-third octave bands:

$$\text{DBSTOI} = \frac{1}{QM} \sum_{m=1}^{M} \sum_{q=1}^{Q} \bar{\rho}_{q,m}. \tag{G.15}$$

This value is influenced by the choice of $\gamma_{k,m}$ and $\tau_{k,m}$. These values are fixed within each one third octave band, so that $\gamma_{k,m} = \gamma_{q,m}$ for $k_1(q) \leq k \leq k_2(q)$. At the same time, the values are fixed across one vector of envelope samples, (G.9). The values are chosen such as to maximize the intermediate correlation, i.e. the predicted intelligibility:

$$\bar{\rho}_{q,m} = \max_{\gamma_{q,m}, \tau_{q,m}} \quad \bar{\rho}_{q,m}(\gamma_{q,m}, \tau_{q,m}). \tag{G.16}$$

In practice, this optimization is carried out as a simple grid search across all combinations of -1 ms $< \tau_{q,m} < +1$ ms in 100 uniform steps and -20 dB $< \gamma_{q,m} < +20$ dB in 40 steps. In addition, we also search the two "better-ear"-options given by the $(\gamma_{q,m}, \tau_{q,m})$-pairs $(+\infty, 0)$ and $(-\infty, 0)$. This corresponds to listening with only one ear and ignoring the other.

The DBSTOI measure outputs a rating of intelligibility on a somewhat arbitrary scale from zero to one. To obtain a prediction of measured intelligibility as a percentage of correctly repeated words, we transform this output with a logistic function [44]:

$$f(d) = \frac{100\%}{1 + e^{ad+b}}, \tag{G.17}$$

where d is the DBSTOI measure, $f(d)$ is the estimated fraction of correctly repeated words and a and b are free parameters, which we fit by maximum likelihood, to provide the best possible predictions [44].

3.2 The DBSTOI measure in spatially diverse conditions

In [44] the DBSTOI measure is evaluated in a range of conditions including different noise types, different configurations of interferers, as well as IBM processing and beamforming. Generally, the measure was found to correlate well with intelligibility in the investigated conditions. However, some indication was also seen of the measure providing inaccurate predictions, when making comparisons between conditions with a point source interferer and conditions with multiple or distributed interferers (e.g. isotropic noise). In this section, we investigate this effect further.

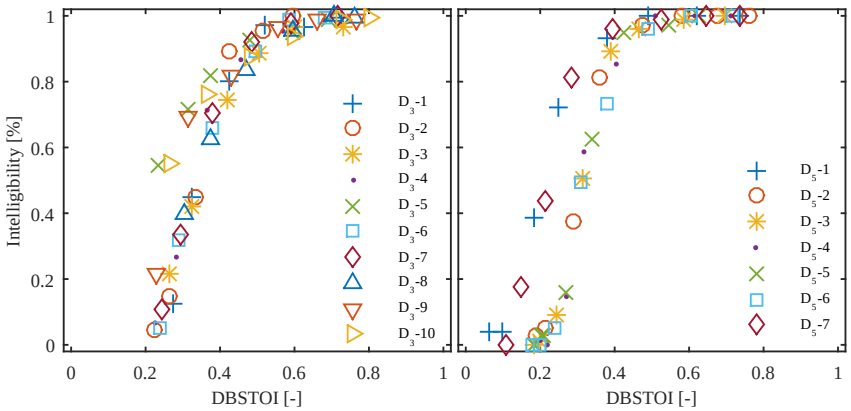


Fig. G.2: Output of the DBSTOI measure plotted against measured intelligibility for datasets D_3 [42] (left) and D_5 (right).

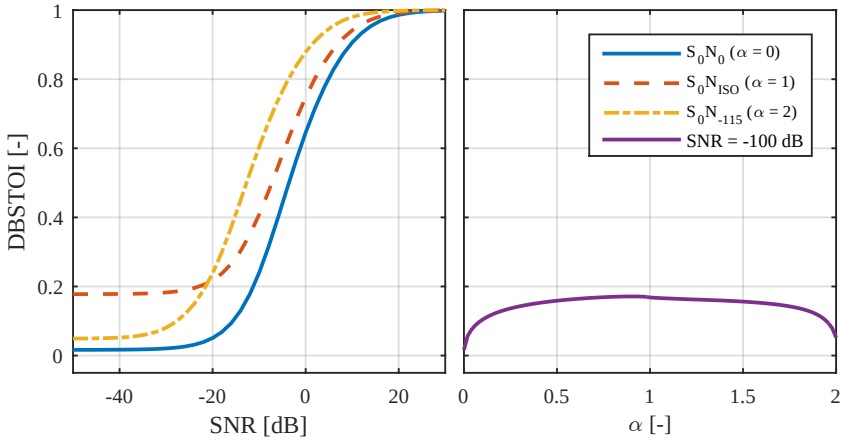


Fig. G.3: The output of the DBSTOI measure at low SNRs. Left: the DBSTOI measure as a function of SNR in three acoustical conditions. Right: the output of the DBSTOI measure at an SNR of -100 dB versus the value of α, as defined by (G.1).

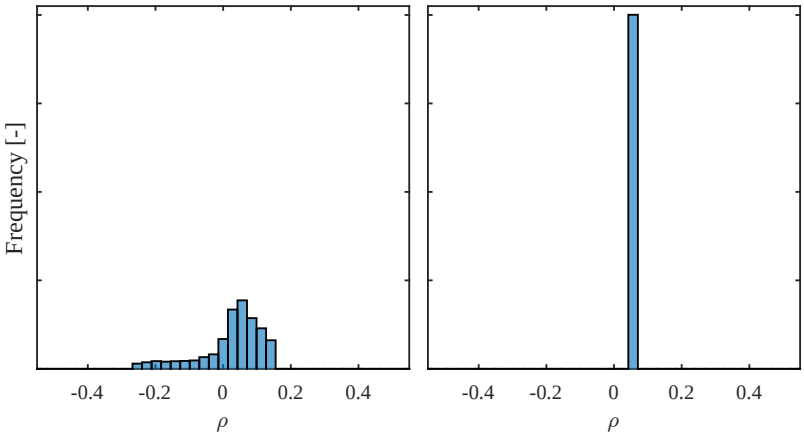


Fig. G.4: Examples of the distribution of intermediate correlation coefficients, (G.12), across the values sampled by the grid search, (G.16), for a random time frame and for band $q = 4$ (with center frequency 300 Hz). Left: all four input signals are independent white noise signals. Right: clean and degraded inputs are independent white noise signals, but the left and right ear signals are identical for both pairs.

Fig. G.2 shows measurements from datasets D_3 and D_5, which contain both distributed and point-like interferers, plotted against the DBSTOI measure of each condition. The left plots shows the discrepancy which was noted in [44]. It appears that the conditions cluster along two distinct lines. The left cluster, which has relatively high measured intelligibility compared with predictions, consists mainly of the conditions with two ISTS interferers. The two conditions with two SSN interferers fall somewhere between the two clusters. The right cluster, containing the remaining conditions, terminates at a DBSTOI measure just above 0.2 and a measured intelligibility of about 0%. This cluster consists of the conditions with more spatially distributed interferers (i.e those with either three point sources or isotropic noise). The right plot of Fig. G.2 shows predictions for dataset D_5. Here, we see the same tendency towards clustering. A left cluster contains the two conditions which exclusively consist of point interferers, while a right cluster contains the conditions where isotropic noise is included. For both datasets, the clusters join together at high levels of intelligibility, and the discrepancy is most pronounced at low intelligibility levels.

To shed further light on this issue, we investigate the behaviour of the DBSTOI measure at very low SNRs. For this purpose, we consider the predictions made for frontal speech masked by the same type of interferer signals used in dataset D_5, as given by (G.1). Fig. G.3 shows the DBSTOI

measure for different SNRs, for a frontal target, masked by noise signals corresponding to different α-values. On the left, the DBSTOI measure is plotted against SNR, for three different α-values. It is immediately apparent that the DBSTOI measure does not converge exactly to 0, at low SNRs, for any of the conditions. The right plot shows the DBSTOI measure versus α at an SNR of -100 dB, i.e. a situation where the output of the DBSTOI measure should ideally be zero. This shows that the DBSTOI measure almost reaches 0 for the conditions with a point source interferer ($\alpha = 0$ and $\alpha = 2$), but settles at a higher value, whenever isotropic noise is present. This is a significant insight into the discrepancies seen in Fig. G.2, as these can be characterized by the DBSTOI measure predicting too high intelligibility at low SNRs, whenever isotropic noise or multiple distributed interferers are present.

To understand this issue further, we look at the internals of the DBSTOI measure. When computing each intermediate correlation coefficient, as described by (G.16), the DBSTOI measure does a grid search of (100 values of τ)$\times$(40 values of γ)+(2 better ear options)$=$ 4002 parameter combinations, with the aim of finding the combination that leads to the highest predicted intelligibility. However, this optimization process is not only influenced by intelligibility, but also by random fluctuations in the signals. To illustrate this point, the left plot of Fig. G.4 shows a histogram of the intermediate correlation, $\bar{\rho}_{q,m}$, for all 4002 possible (τ, γ)-combinations for one frame, when the four input signals are uncorrelated white Gaussian noise. Since the clean and degraded signals are entirely uncorrelated in this scenario, the intermediate correlation should be 0 on average. However, the figure shows that a considerable variation occurs for the different EC parameters. When always picking the highest correlation from such distributions (i.e. what is done in (G.16)), a considerable upwards bias is introduced. The right plot of Fig. G.4 shows the distribution arising from a similar procedure, but when the left and right signals are identical white Gaussian noise signals, while the clean and degraded signals are still uncorrelated. Here, it is seen that the intermediate correlation is the same for all sampled EC-parameters. In [44] it was shown that the DBSTOI measure reduces to the monaural STOI measure under such conditions. Thereby, no variation can be introduced in the EC stage. This explains why the upwards bias is not present, when highly correlated signals are presented to the left and the right ear, e.g. when a single interferer is located in front of the listener together with the target.

The mechanism described above presents a fundamental issue with the DBSTOI measure: a signal dependent upward bias is introduced at low SNRs. This can be more or less disruptive, depending on the application of the measure. When using it to compare predicted intelligibility between acoustically similar conditions, the issue is not severe, as the bias will be present to the same extent in all predictions. However, if using the measure

to compare widely different acoustical situations, the bias may be of different magnitude in the different conditions and thus corrupt the comparability of predictions. The issue may also be disruptive when comparing different types of processing (e.g. speech enhancement or beamforming) as these may modify the binaural cues of the signal differently, thus effectively changing the acoustical scene conveyed by the binaural signal. Since such comparisons of different acoustical environments and different processing types are central use-cases of the DBSTOI measure, this issue is highly undesirable. We therefore propose a slight modification to the DBSTOI measure, which greatly diminishes the bias.

3.3 The MBSTOI measure

To approach the problematic bias, we consider a subset of the domain in which the DBSTOI measure is meant to be applied. We derive analytically a bias compensation strategy for this domain. In Section 4, we demonstrate that the strategy works well outside of this domain. Let us, therefore, consider the situations where the degraded signal can be written as a sum of the clean signal and a noise/distortion component:

$$\begin{aligned} y_l(t) &= x_l(t) + n_l(t), \\ y_r(t) &= x_r(t) + n_r(t). \end{aligned} \tag{G.18}$$

Here, $n_l(t)$ and $n_r(t)$ may be additive noise or distortion, but are assumed to be uncorrelated with the clean reference signals $x_l(t)$ and $x_r(t)$. In this situation, the degraded speech power envelope (defined by (G.7)) can be written as:

$$\begin{aligned} Y_{q,m} &= \sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m} + \hat{n}_{k,m}|^2 \\ &= \underbrace{\sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m}|^2}_{X_{q,m}} + \underbrace{\sum_{k=k_1(q)}^{k_2(q)} |\hat{n}_{k,m}|^2 + \sum_{k=k_1(q)}^{k_2(q)} 2\mathrm{Re}(\hat{x}_{k,m}^* \hat{n}_{k,m})}_{R_{q,m}}, \end{aligned} \tag{G.19}$$

where the first term is recognized as the clean signal power envelope, $X_{q,m}$, and two following terms are collectively referred to as $R_{q,m}$. We now define an envelope SNR:

$$\mathrm{SNR}_{q,m} \triangleq \frac{\sum_{k=k_1(q)}^{k_2(q)} |\hat{x}_{k,m}|^2}{\sum_{k=k_1(q)}^{k_2(q)} |\hat{n}_{k,m}|^2}. \tag{G.20}$$

Considering a situation with very low envelope SNR, i.e. $\mathrm{SNR}_{q,m} \to 0$, we have[3]:

$$Y_{q,m}\Big|_{\mathrm{SNR}_{q,m}\ll 1} \approx R_{q,m}\Big|_{\mathrm{SNR}_{q,m}\ll 1} \approx \sum_{k=k_1(q)}^{k_2(q)} |\hat{n}_{k,m}|^2. \tag{G.22}$$

The above approximations become exact for SNRs tending toward negative infinity.

Now, define:

$$\mathbf{r}_{q,m} = [R_{q,m-N+1}, \ldots, R_{q,m}]^\intercal - \mathbf{1} \sum_{m'=m-N+1}^{m} \frac{R_{q,m'}}{N}. \tag{G.23}$$

From this point, we drop the frame and frequency band indexes, q and m, from $\mathbf{x}_{q,m}$, $\mathbf{y}_{q,m}$ and $\mathbf{r}_{q,m}$ for notational convenience. Then (G.19) implies $\mathbf{y} = \mathbf{x} + \mathbf{r}$. Inserting this into (G.12), we obtain:

$$\begin{aligned}\bar{\rho}_{q,m} &= \frac{E_\Delta\left[\mathbf{x}^\intercal\mathbf{x}\right] + E_\Delta\left[\mathbf{x}^\intercal\mathbf{r}\right]}{\sqrt{E_\Delta\left[\mathbf{x}^\intercal\mathbf{x}\right] E_\Delta\left[\mathbf{y}^\intercal\mathbf{y}\right]}} \\ &= \sqrt{\frac{E_\Delta\left[\mathbf{x}^\intercal\mathbf{x}\right]}{E_\Delta\left[\mathbf{y}^\intercal\mathbf{y}\right]}} + \frac{E_\Delta\left[\mathbf{x}^\intercal\mathbf{r}\right]}{\sqrt{E_\Delta\left[\mathbf{x}^\intercal\mathbf{x}\right] E_\Delta\left[\mathbf{y}^\intercal\mathbf{y}\right]}}.\end{aligned} \tag{G.24}$$

Here, $E_\Delta[\cdot]$ is the expectation across the jitter from the EC stage. We remind the reader that (G.19), (G.22) and (G.24) are functions of the EC parameters, τ and γ. Clearly, (G.22) holds independently of these values, except for the special case where the EC stage would provide an infinite improvement in SNR (i.e. in the hypothetical situation where the EC stage is able to completely cancel the noise). The jitter effectively prevents this from happening in practice [32]. For low SNRs, we thus have $E_\Delta\left[\mathbf{y}^\intercal\mathbf{y}\right] \approx E_\Delta\left[\mathbf{r}^\intercal\mathbf{r}\right]$ and $E_\Delta\left[\mathbf{y}^\intercal\mathbf{y}\right] \gg E_\Delta\left[\mathbf{x}^\intercal\mathbf{x}\right]$. Inserting this into (G.24), we get:

$$\bar{\rho}\Big|_{\mathrm{SNR}_{q,m}\ll 1} \approx \quad 0 \quad + \quad \frac{E_\Delta\left[\mathbf{x}^\intercal\mathbf{r}\right]}{\sqrt{E_\Delta\left[\mathbf{x}^\intercal\mathbf{x}\right] E_\Delta\left[\mathbf{r}^\intercal\mathbf{r}\right]}}. \tag{G.25}$$

The first term approaches zero, and the second term is a short-time estimate of the correlation between the signals $X_{q,m'}$ and $R_{q,m'}$ for $m' = m - N +$

[3]To see the result from (G.22), decompose the clean signals, $x_l(t)$ and $x_r(t)$, as $x_l(t) = c\phi_l(t)$ and $x_r(t) = c\phi_r(t)$, where $c > 0$ and $\phi_l(t)$ and $\phi_r(t)$ are arbitrary real signals. Inserting this in (G.19) yields:

$$Y_{q,m} = c^2 \sum_{k=k_1(q)}^{k_2(q)} |\hat{\phi}_{k,m}|^2 + \sum_{k=k_1(q)}^{k_2(q)} |\hat{n}_{k,m}|^2 + c \sum_{k=k_1(q)}^{k_2(q)} 2\mathrm{Re}(\hat{\phi}^*_{k,m}\hat{n}_{k,m}). \tag{G.21}$$

It is evident that the value of c directly controls $\mathrm{SNR}_{q,m}$. The first term diminishes for small c, i.e. for low $\mathrm{SNR}_{q,m}$, leaving only $R_{q,m}$. As $c \to 0$, only the second term remains.

$1, \ldots, m$. This second term is not in general zero, and is furthermore dependent on the EC parameters. We can conclude that the bias at low SNRs must stem from the fact that the EC stage parameters, γ and τ, are found such as to maximize the second term in (G.25). We therefore propose a modified objective function for finding EC stage parameters, in which only the first term of (G.24) is included. In other words, rather than finding the EC parameters, $(\gamma_{q,m}, \tau_{q,m})$, which maximize (G.24), we find instead the EC parameters, $(\gamma_{q,m}, \tau_{q,m})$, which maximize:

$$g = \sqrt{\frac{E_\Delta\left[\mathbf{x}^\intercal \mathbf{x}\right]}{E_\Delta\left[\mathbf{y}^\intercal \mathbf{y}\right]}}. \tag{G.26}$$

Note that by doing so, we ensure that in situations where the true intelligibility is certainly going to be 0% (additive uncorrelated noise at sufficiently low SNR), the predictor output is also going to be zero. Furthermore, it should be noted that in high SNR situations, i.e. where $E_\Delta\left[\mathbf{x}^\intercal \mathbf{x}\right] \approx E_\Delta\left[\mathbf{y}^\intercal \mathbf{y}\right]$, we have $g = 1$. We only use (G.26) to find the values of γ and τ. The found parameters are then used to evaluate the intermediate correlation, (G.12), as in the DBSTOI measure. We refer to the modified measure as the MBSTOI measure.

In these derivations, we assumed a very simple signal model, (G.18), which is not strictly valid for all situations, in which the MBSTOI measure may find use, i.e. when the degraded signal is a processed version of the noisy speech signal, and this processing cannot be modeled as an uncorrelated additive component. However, an argument, similar to the one carried out above, leads to the same method, if the noisy speech has been processed, and this processing is 1) identical on both ears, 2) linear, 3) slowly time varying (with respect to the duration of N DFT windows), and 4) approximately constant within each one-third octave band[4]. We hypothesize, furthermore, that the MBSTOI measure works well in an even broader context. This is experimentally verified in Section 4.

3.4 The beMBSTOI measure

We also define a better ear version of the MBSTOI measure. This is easily obtained by performing the optimization across a search grid of only two $(\gamma_{q,m}, \tau_{q,m})$-pairs: $(+\infty, 0)$ and $(-\infty, 0)$. Mathematically, this corresponds completely to disabling the EC stage, computing the (modified) STOI measure for each ear separately, and maintaining only the larger result. This measure allows us to investigate the contribution of binaural unmasking to predicted speech intelligibility. We refer to it as the beMBSTOI measure.

[4] In this case, the processing amounts approximately to a constant factor, $c_{q,m}^2$, multiplied onto (G.19).

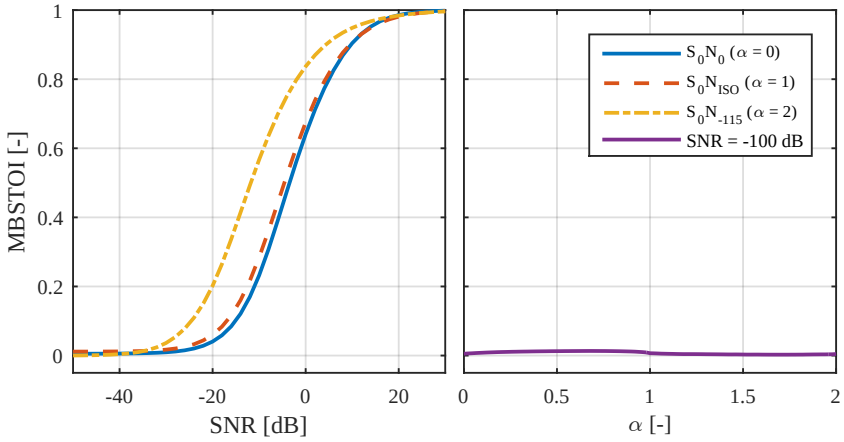


Fig. G.5: The output of the MBSTOI measure at low SNRs. Left: the MBSTOI measure as a function of SNR in three acoustical conditions. Right: the output of the MBSTOI measure at an SNR of -100 dB versus the value of α, as defined by (G.1).

4 Results and Discussion

We investigate the performance of the proposed MBSTOI measure and other measures in terms of 1) bias characteristics, 2) general prediction performance, and 3) prediction of binaural advantage. We furthermore, discuss the relationship between the different evaluated SIP algorithms, as well as the extent to which the results generalize to conditions beyond those tested here.

4.1 Bias in the MBSTOI measure

The bias problem, encountered in the DBSTOI measure, is most directly apparent from Fig. G.3. Fig. G.5 therefore shows the same plot, using the same input signals, but for the MBSTOI measure. From this figure, it appears that almost no upwards bias is present in the MBSTOI measure. The right plot on the figure shows that a slight upwards bias is still present for α-values just under 1. However, the magnitude of this effect is so small that it has no practical impact on prediction performance.

Fig. G.5 indicates that upwards bias, in conditions with spatially distributed interferers, is not a pronounced issue in the MBSTOI measure, but it does not say anything about how the proposed modification impacts prediction performance. This is investigated in the following two sections.

4.2 General Prediction Performance

In this section we investigate prediction performance for each of the five datasets and for four different SIP algorithms: the DBSTOI measure [44], the MBSTOI measure, the beMBSTOI measure, and the B-sEPSM [45].

The B-sEPSM, proposed in [45], is a binaural extension of the mr-sEPSM, which handles binaural and non-linearly processed stimuli. The B-sEPSM bases predictions on a degraded speech signal and a clean *noise* signal, i.e. noise in the absence of speech and processing. An implementation of the measure has kindly been made available by the authors [55][5]. The B-sEPSM outputs a number, B-SNR$_{\text{env}}$, which is transformed to a sensitivity index [45]:

$$d' = k(\text{B-SNR}_{\text{env}})^q, \tag{G.27}$$

and further to a prediction of the probability that a subject answers correctly [45]:

$$P_{correct} = \Phi\left(\frac{d' - \mu_N}{\sqrt{\sigma_S^2 + \sigma_N^2}}\right), \tag{G.28}$$

where $\Phi(\cdot)$ is a cumulative normal distribution, k, q, μ_N, and σ_N^2 are constants determined by the speech material and test setup, while σ_S^2 is fitted to the data at hand. In this work, we use the B-sEPSM slightly differently to make it more comparable with the other methods. We first transform the output, B-SNR$_{\text{env}}$, by (G.27) using $k = \sqrt{1.2}$ and $q = 0.5$. These values were suggested in [24] for making predictions together with the Dantale II speech material, which was used in collecting the five datasets considered in this study. We then transform the resulting sensitivity index, d', by a logistic function, (G.17), and fit the parameters, a, and b to the available data, exactly as is done for the other measures discussed in this paper. It should be noted that the logistic function is very similar to the cumulative normal distribution, and that the main difference therefore lies in the fact that we fit two parameters (a, b) to the available data rather than one (σ_S^2).[6]

Fig. G.6 shows the predictions of the four measures for each of the five datasets. It was not possible to obtain clean noise signals for dataset D_4, and B-sEPSM predictions for this dataset are therefore missing. A single logistic function was fitted for each measure (i.e. a and b are constant across each row in the figure).

[5]The implementation is part of the Python Auditory Modelling Toolbox. See `http://pambox.org/`.

[6]The B-sEPSM only processes frequency channels which exceed the absolute threshold of hearing [45]. The datasets used in this study were collected at normal conversational level, and thus not near the absolute threshold of hearing. We therefore ensured that the signals provided to the B-sEPSM were always powerful enough for all frequency channels to be processed.

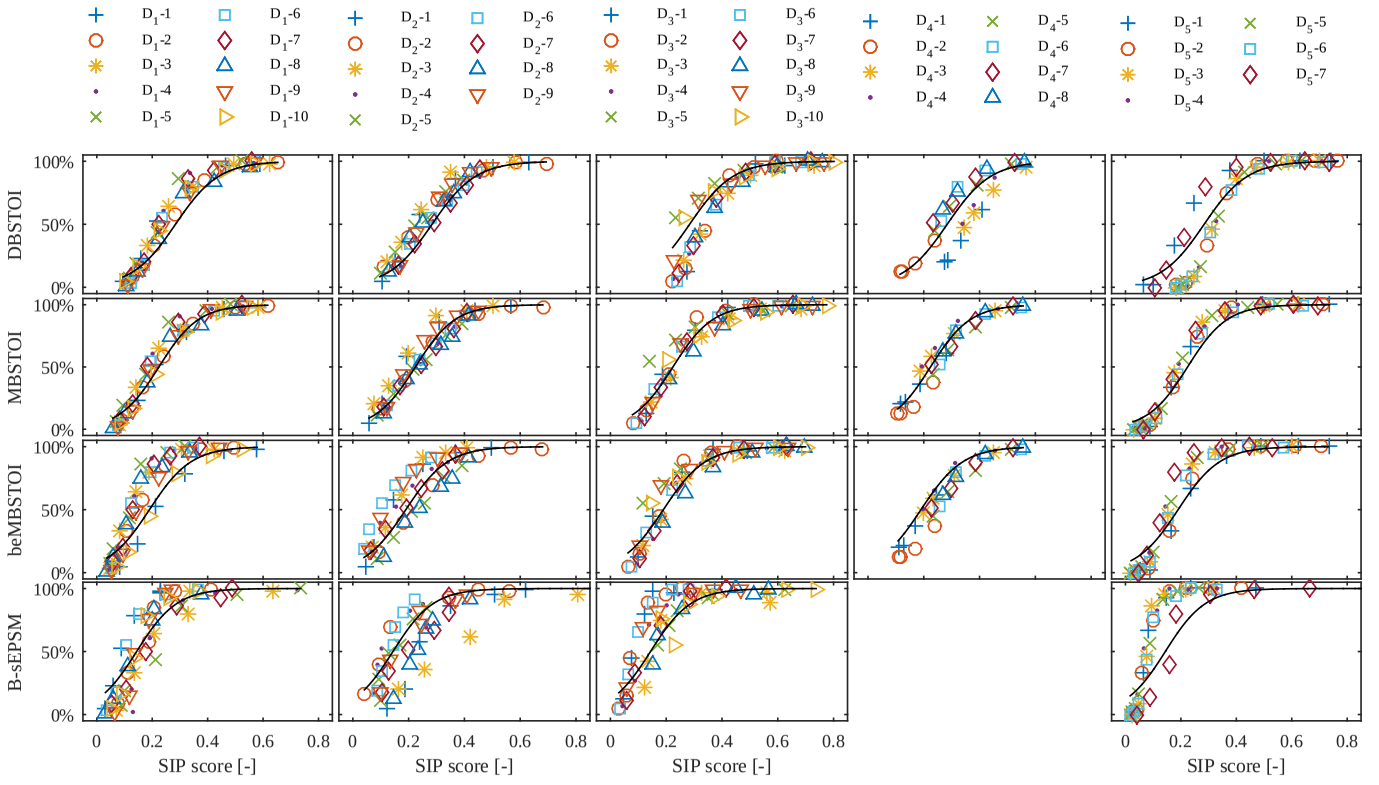


Fig. G.6: The outputs of the four studied intelligibility predictors (rows) plotted against the measured intelligibility of the five studied datasets (columns). Details about the conditions are given in Table G.1. For the B-sEPSM the plot shows d', (G.27), divided by a factor of 50, such as to make it fit on the same axis as the other predictors. Predictions for the D_4-data are missing for the B-sEPSM, because we were unable to obtain clean noise signals for this dataset. The black curves are logistic functions, (G.17). Only one logistic function was fitted for each predictor (row).

The DBSTOI measure (Fig. G.6, row 1) makes accurate predictions for the datasets with point source interferers, D_1 and D_2, while for the remaining datasets, D_3, D_4, and D_5, predictions fall in two clusters depending on whether the interferers are point-like or spatially distributed. The MBSTOI measure (Fig. G.6, row 2) makes almost identical predictions for datasets D_1 and D_2. For the three remaining datasets, D_3, D_4, and D_5, the bias problem appears to be alleviated. Some deviation is, however, still seen for conditions $D_3 - 5$ and $D_3 - 10$, which contain two ISTS interferers. This corresponds to unintelligible two-talker babble, which is strongly fluctuating. The STOI measure is known to underestimate intelligibility for such interferers [26], and this is most likely the cause of observed deviations. It therefore appears that the proposed modification to the DBSTOI measure serves its intended purpose, without otherwise impairing prediction performance considerably. The tested conditions include a substantial range of combinations of different acoustical setups, different interferer types, noise reduction and beamforming. It therefore spans many of the types of conditions in which the MBSTOI measure could realistically be applied.

The third row in Fig. G.6 shows predictions by the beMBSTOI measure, which does not have the EC stage, but works purely in a better-ear fashion. For datasets D_1 and D_2, it is very evident that the beMBSTOI measure is not accounting for binaural unmasking, i.e. predictions are too low for conditions, where this provides an advantage. This is especially evident for dataset D_1, where intelligibility is underestimated in all conditions except the two where the interferer is located directly in front or directly behind. In conditions such as those in datasets D_1 and D_2, a simple better-ear approach is thus not sufficient to predict binaural advantage. A similar effect can be seen for dataset D_5. However, for datasets D_3 and D_4, predictions do not appear to be significantly impaired. These datasets cover conditions which are complicated in terms of both spatially distributed interferers and processing. At the same time they do not contain conditions with a single point-like interferer located away from the frontal position. The absence of such "pure" conditions, with large potential for binaural unmasking, may explain why predictions are reasonably good without the EC stage.

The last row of Fig. G.6 shows predictions of the B-sEPSM. These follow the trend of the measurements, but are less accurate than the predictions of the MBSTOI measure. From dataset D_1, it appears that binaural advantage is overestimated. At the same time, from dataset D_2, it appears that the effect of IBM processing may also be overestimated. The B-sEPSM uses an EC stage similar to that of the MBSTOI measure, but the underlying SIP algorithm, the mr-sEPSM [21], is very different from the STOI measure. It is designed to take modulation masking into account, and therefore performs well in conditions with modulated interferers [21, 56]. The datasets considered in this paper lack such conditions.

Table G.2: Summary of prediction performance ([RMSE] / [Kendall's Tau]).

	D_1	D_2	D_3	D_4	D_5
DBSTOI	10.80% / 0.88	8.73% / **0.87**	12.87% / 0.84	13.54% / 0.70	15.39% / 0.81
MBSTOI	**7.34%** / **0.89**	**6.92%** / 0.86	**7.07%** / 0.84	**6.30%** / **0.91**	**7.69%** / **0.89**
beMBSTOI	16.57% / 0.80	13.06% / 0.73	8.49% / **0.85**	10.71% / 0.89	11.27% / 0.88
B-sEPSM	15.12% / 0.75	17.82% / 0.67	16.12% / 0.69	— / —	23.48% / 0.82

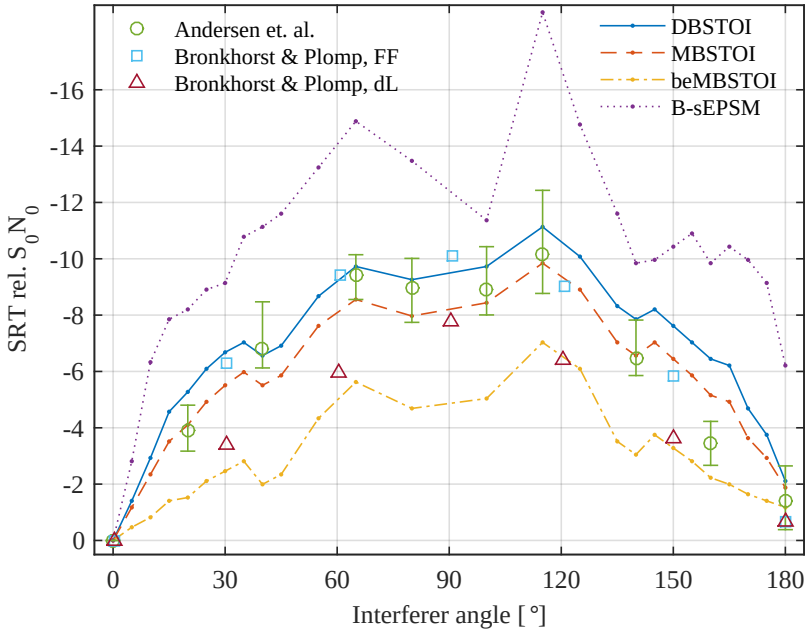


Fig. G.7: Binaural advantage as measured in two independent studies, and predicted by four SIP algorithms. The error bars on the "Andersen et al."-data show quartiles (the binaural advantage was computed separately for the 10 subjects).

The overall performance of the four methods is summarized in Table G.2, in terms of Root-Mean-Square Error (RMSE) and Kendall's Tau [57]. These metrics support the previous observations. Specifically, the minimum RMSE is obtained by the MBSTOI measure for all five datasets. We remark that Kendall's Tau is independent of one-to-one mappings, and that the fitted logistic mapping therefore has no effect on this metric. This also means that performance is independent of whether the B-sEPSM framework is used as described in [45], or in the slightly different manner in which we use it here.

4.3 Prediction of Binaural Advantage

In this section we evaluate the performance of the four methods in terms of predicting binaural advantage. We therefore focus specifically on the results of dataset D_1, which contains conditions with a frontal target and a single interferer at different azimuths, i.e. conditions where binaural advantage is expected to be high.

The SRT was estimated for each of these conditions, by maximum-likelihood-fitting a logistic function to measured intelligibility as a function of SNR, and computing the intersection with 50% intelligibility [44]. We define the binaural advantage of each condition of dataset D_1 to be the difference in SRT between the particular condition and the condition with an interferer azimuth of 0°, i.e. condition D_1-6 (see Table G.1). The binaural advantage is plotted in Fig. G.7 ("Andersen et al."). Similar measurements were carried out by [31] and these are also included ("Bronkhorst & Plomp, FF"). This study also measured binaural advantage with artificially processed stimuli, where no binaural unmasking is possible (see [31] for details). For this type of stimuli, no additional performance is predicted by using an EC stage relative to better-ear processing as used in the beMBSTOI measure. These results are also shown in Fig. G.7 ("Bronkhorst & Plomp, dL").

To predict binaural advantage using the studied predictors, calibration values were generated by evaluating each predictor at the SRT in condition D_1-6 (which has the target and interferer co-located at the front). The calibration value for each predictor was then assumed to always correspond to the SRT. Hence, the SRT in another condition can be predicted by adaptively adjusting the input SNR to a predictor, until the predictor output is equal to the calibration value. The binaural advantage, predicted with the four studied predictors in this way, is shown in Fig. G.7.

From Fig. G.7 we first note that the DBSTOI measure predicts binaural advantage accurately in these conditions, as was also concluded in [44]. The DBSTOI measure predicts binaural advantage to increase as the interferer is moved away from 0°. Peaks are located both below and above 90°. This well-known phenomenon, which is most likely caused by the acoustics of the human head [27], is also seen in the "Andersen et al."-data.

The MBSTOI measure predicts binaural advantage to be up to 1.5 dB smaller than the DBSTOI measure does. This could be due to the MBSTOI measure also removing some bias in these conditions. Fig. G.3 indicates a slight bias also in the S_0N_{-115}-condition (the DBSTOI measure does not converge to 0 for low SNRs). While this happens to decrease prediction performance of the MBSTOI measure relative to the DBSTOI measure, the MBSTOI-predictions are still well in line with the measurements. This, furthermore, should be seen in the light of the considerable variability inherent to such measurements (as seen from the error bars on Fig. G.7).

As expected, the beMBSTOI measure predicts binaural advantage to be much smaller than the other methods, because it does not include the prediction of binaural unmasking. It, however, reveals the predicted better ear advantage in isolation. Similarly, the difference in predictions by the MBSTOI and beMBSTOI measures show the amount of binaural unmasking included in predictions by the MBSTOI measure. This is seen to be around 2 dB when the interferer is not near 0° or 180°. Fig. G.7 also shows results from [31], of measured binaural advantage ("Bronkhorst & Plomp, FF") and of binaural advantage, measured using modified stimuli that does not allow for binaural unmasking ("Bronkhorst & Plomp, dL"). The vertical distance between each corresponding pair of points from "Bronkhorst & Plomp, dL" and "Bronkhorst & Plomp, FF" amounts to an empirical observation of binaural unmasking. It is interesting to note that the distance between these measurements correspond well with the distance between prediction by the MBSTOI and beMBSTOI measures. Moreover, the beMBSTOI measure predicts the data points of "Bronkhorst & Plomp, dL" quite well. This indicates that the MBSTOI measure correctly estimates the separate contributions of better-ear advantage and binaural unmasking.

The B-sEPSM overestimates binaural advantage considerably. This result is somewhat surprising, as the B-sEPSM also uses an EC stage, similar to that of the MBSTOI measure, to predict binaural advantage. However, the fact that it uses a degraded signal and a clean *noise* signal, could possibly alter the operation of the EC stage. Furthermore, as mentioned, the underlying SIP algorithm, the mr-sEPSM, is significantly different from the STOI measure. The result is consistent with observations of [45], which, however, also indicate that the B-sEPSM performs well in more complicated binaural conditions with reverberation and fluctuating interferers.

5 Conclusions

In this paper we have carried out a detailed investigation of the Deterministic Binaural Short-Time Objective Intelligibility (DBSTOI) measure in acoustical conditions with distributed or multiple interferers. We showed that, in such conditions at low Signal to Noise Ratios (SNRs), the DBSTOI measure suffers from an upward bias. The reason for this bias was studied in detail, resulting in a modified algorithm which does not suffer from this issue: the Modified BSTOI (MBSTOI) measure. Across five different datasets, the MBSTOI measure performed similarly to the DBSTOI measure in conditions with point-like interferers, and outperformed the DBSTOI measure in conditions with multiple or spatially distributed interferers. Within the scope of the five datasets used for testing, the MBSTOI measure essentially removed the issues relating to bias, with no apparent side effects.

6 Acknowledgments

This work was supported by the Oticon Foundation and the Danish Innovation Foundation. We thank Adam Kuklasiński for providing dataset D_4.

References

[1] H. Fletcher, J. C. Steinberg, Articulation testing methods, Bell System Technical Journal 8 (4) (1929) 806–854.

[2] N. R. French, J. C. Steinberg, Factors governing the intelligibility of speech sounds, J. Acoust. Soc. Am. 19 (1) (1947) 90–119.

[3] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An algorithm for inteligibility prediction of time-frequency weighted noisy speech, IEEE Tran. on Audio, Speech and Language Processing 19 (7) (2011) 2125–2136.

[4] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, S. Scollie, Objective quality and intelligibility prediction for users of assistive listening devices, IEEE Signal Processing Magazine 32 (2) (2015) 114–124.

[5] C. Ludvigsen, C. Elberling, G. Keidser, Evaluation of a noise reduction method – comparison between observed scores and scores predicted from STI, Scand. Audiol. Suppl. 38 (1993) 50–55.

[6] R. L. Goldsworthy, J. E. Greenberg, Analysis of speech-based speech transmission index methods with implications for nonlinear operations, J. Acoust. Soc. Am. 116 (6) (2004) 3679–3689.

[7] J. B. Boldt, D. P. W. Ellis, A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation, in: Proceedings of the 17th European Signal Processing Conference (EUSIPCO), EURASIP, Glasgow, Scotland, 2009, pp. 1849–1853.

[8] J. Jensen, C. H. Taal, Speech intelligibility prediction based on mutual information, IEEE Tran. on Audio, Speech and Language Processing 22 (2) (2014) 430–440.

[9] S. van Wijngaarden, The speech transmission index after four decades of development, Acoustics Australia 40 (2) (2012) 134–138.

[10] Sound System Equipment. Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index, Standard, International Organization for Standardization, Geneva, Switzerland (2011).

[11] S. Jørgensen, J. Cubick, T. Dau, Speech intelligibility evaluation for mobile phones, Acta Acustica United with Acustica 101 (2015) 1016–1025.

[12] J. B. Allen, The articulation index is a shannon channel capacity, in: D. Pressnitzer, A. de Cheveigne, S. McAdams, L. Collet (Eds.), Auditory Signal Processing: Physiology, Psychoacoustics and Models, Springer Verlag, 2004, pp. 314–320.

[13] Methods for Calculation of the Speech Intelligibility Index, Standard, American National Standards Institute, New York, United States (1997).

[14] Y. Hu, P. C. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Tran. on Audio, Speech and Language Processing 16 (1) (2008) 229–238.

[15] J. Ma, Y. Hu, P. C. Loizou, Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions, J. Acoust. Soc. Am. 125 (5) (2009) 3387–3405.

[16] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech, J. Acoust. Soc. Am. 130 (5) (2011) 3013–3027.

[17] K. Smeds, A. Leijon, F. Wolters, A. Hammarstedt, S. Båsjö, S. Hertzman, Comparison of predictive measures of speech recognition after noise reduction processing, J. Acoust. Soc. Am. 136 (3) (2014) 1363—-1374.

[18] G. A. Miller, The masking of speech, Psychological Bulletin 44 (2) (1947) 105–129.

[19] K. S. Rhebergen, N. J. Versfeld, A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners, J. Acoust. Soc. Am. 117 (4) (2005) 2181–2192.

[20] K. S. Rhebergen, N. J. Versfeld, W. A. Dreschler, Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise, J. Acoust. Soc. Am. 120 (6) (2006) 3988–3997.

[21] S. Jørgensen, S. D. Ewert, T. Dau, A multi-resolution envelope-power based model for speech intelligibility, J. Acoust. Soc. Am. 134 (1) (2013) 436–446.

[22] M. Cooke, Glimpsing model of speech perception, J. Acoust. Soc. Am. 119 (3) (2006) 1562–1573.

[23] J. M. Kates, K. H. Arehart, Coherence and the speech intelligibility index, J. Acoust. Soc. Am. 117 (4) (2005) 2224–2237.

[24] S. Jørgensen, T. Dau, Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing, J. Acoust. Soc. Am. 130 (3) (2011) 1475–1487.

[25] S. D. Ewert, T. Dau, Characterizing frequency selectivity for envelope fluctuations, J. Acoust. Soc. Am. 108 (3) (2000) 1181–1196.

[26] J. Jensen, C. Taal, An algorithm for predicting the intelligibility of speech masked by modulated noise maskers, IEEE Tran. on Audio, Speech and Language Processing 24 (11) (2016) 2009–2022.

[27] A. W. Bronkhorst, The cocktail party phenomenon: A review on speech intelligibility in multiple-talker conditions, Acta Acustica United with Acustica 86 (1) (2000) 117–128.

[28] S. Doclo, T. Klasen, V. den Bogaert, M. Moonen, J. Wouters, Theoretical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions, in: Int. Workshop Acoust. Echo Noise Control (IWAENC), Paris, France, 2006.

[29] T. Klasen, T. V. den Bogaert, M. Moonen, J. Wouters, Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues, IEEE Tran. on Audio, Speech and Language Processing 55 (4) (2007) 1579–1585.

[30] B. C. Moore, An introduction to the psychology of hearing, sixth Edition, Brill, 2013.

[31] A. W. Bronkhorst, R. Plomp, The effect of head-induced interaural time and level differences on speech intelligibility in noise, J. Acoust. Soc. Am. 83 (4) (1988) 1508–1516.

[32] N. I. Durlach, Equalization and cancellation theory of binaural masking-level differences, J. Acoust. Soc. Am. 35 (8) (1963) 1206–1218.

[33] N. I. Durlach, Binaural signal detection: Equalization and cancellation theory, in: J. V. Tobias (Ed.), Foundations of Modern Auditory Theory Volume II, Academic Press, New York, 1972, pp. 371–462.

[34] R. Beutelmann, T. Brand, Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners, J. Acoust. Soc. Am. 120 (1) (2006) 331–342.

[35] R. Beutelmann, T. Brand, B. Kollmeier, Revision, extension and evaluation of a binaural speech intelligibility model, J. Acoust. Soc. Am. 127 (4) (2010) 2479–2497.

[36] R. Wan, N. I. Durlach, H. S. Colburn, Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers, J. Acoust. Soc. Am. 128 (6) (2010) 3678–3690.

[37] R. Wan, N. I. Durlach, S. Colburn, Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers, J. Acoust. Soc. Am. 136 (2) (2014) 768–776.

[38] M. Lavandier, J. F. Culling, Prediction of binaural speech intelligibility against noise in rooms, J. Acoust. Soc. Am. 127 (1) (2010) 387–399.

[39] S. Jelfs, J. F. Culling, M. Lavandier, Revision and validation of a binaural model for speech intelligibility in noise, Hearing Research 275 (1–2) (2011) 96–104.

[40] M. Lavandier, S. Jelfs, J. F. Culling, A. J. Watkins, A. P. Raimond, S. J. Makin, Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources, J. Acoust. Soc. Am. 131 (1) (2012) 218–231.

[41] J. F. Culling, M. L. Hawley, R. Y. Litovsky, Erratum: The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources [J. Acoust. Soc. Am. 116, 1057 (2004)], J. Acoust. Soc. Am. 118 (1) (2005) 552.

[42] A. H. Andersen, J. M. de Haan, Z.-H. Tan, J. Jensen, A binaural short time objective intelligibility measure for noisy and enhanced speech, in: Proc. Interspeech, Dresden, Germany, 2015, pp. 2563–2567.

[43] A. H. Andersen, J. M. de Haan, Z.-H. Tan, J. Jensen, A method for predicting the intelligibility of noisy and non-linearly enhanced binaural speech, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Shanghai, China, 2016, pp. 4995–4999.

[44] A. H. Andersen, J. M. de Haan, Z.-H. Tan, J. Jensen, Predicting the intelligibility of noisy and non-linearly processed binaural speech, IEEE Tran. on Audio, Speech and Language Processing 24 (11) (2016) 1908–1920.

[45] A. Chabot-Leclerc, E. N. MacDonald, T. Dau, Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain, J. Acoust. Soc. Am. 140 (1) (2016) 192–205.

[46] A. Kuklasiński, S. Doclo, S. H. Jensen, J. Jensen, Maximum likelihood PSD estimation for speech enhancement in reverberation and noise, IEEE Tran. on Audio, Speech and Language Processing 24 (9) (2016) 1599–1612.

[47] S. Braun, E. A. P. Habets, Dereverberation in noisy environments using reference signals and a maximum likelihood estimator, in: The European Signal Processing Conference (EUSIPCO), 2013.

[48] V. R. Algazi, R. O. Duda, D. M. Thompson, C. Avendano, The CIPIC HRTF database, in: Proccedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, IEEE, New Paltz, New York, 2001, pp. 99–102.

[49] K. Wagener, J. L. Josvassen, R. Ardenkjær, Design, optimization and evaluation of a Danish sentence test in noise, International Journal of Audiology 42 (1) (2003) 10–17.

[50] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, D. Wang, Role of mask pattern in intelligibility of ideal binary-masked noisy speech, J. Acoust. Soc. Am. 126 (3) (2009) 1415–1426.

[51] M. Vestergaard, The eriksholm cd 01: Speech signals in various acoustical environments (1998).

[52] E. R. Pedersen, P. M. Juhl, User-operated speech in noise test: Implementation and comparison with a traditional test, International Journal of Audiology 53 (5) (2014) 336–344.

[53] I. Holube, S. Fredelake, M. Vlaming, B. Kollmeier, Development and analysis of an international speech test signal (ISTS), International Journal of Audiology 49 (12) (2010) 891–903.

[54] H. vom Hövel, Zur bedeutung der übertragungseigenschaften des aussenohrs sowie des binauralen hörsystems bei gestörter sprachübertragung, Ph.D. thesis, Rheinisch-Westfälische Technische Hochschule Aachen (1984).

[55] A. Chabot-Leclerc, PAMBOX: A python auditory modeling toolbox, in: The 7th Speech in Noise Workshop, Copenhagen, Denmark, 2015.

[56] Helia-Relaño-Iborra, T. May, J. Zaar, C. Scheidiger, T. Dau, Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain, J. Acoust. Soc. Am. 140 (4) (2016) 2670–2679.

[57] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1/2) (1938) 81–93.

Paper H

Non-Intrusive Speech Intelligibility Prediction using Convolutional Neural Networks

Asger Heidemann Andersen
Jan Mark de Haan
Zheng-Hua Tan
Jesper Jensen

Submitted to
IEEE/ACM Transactions on Audio, Speech, and Language Processing.

© 2017 IEEE
The layout has been revised.

Abstract

Speech Intelligibility Prediction (SIP) algorithms are becoming popular tools within the development and operation of speech processing devices and algorithms. However, many SIP algorithms require knowledge of the underlying clean speech; a signal which is often not available in real-world applications. This has recently led to increased interest in non-intrusive SIP algorithms, which do not require clean speech to make predictions. In this paper we investigate the use of Convolutional Neural Networks (CNNs) for non-intrusive SIP. To do so, we utilize a CNN architecture which shows similarities to existing SIP algorithms, in terms of computational structure, and which allows for easy and meaningful visualization and interpretation of trained weights. We evaluate this architecture by using a large dataset obtained by combining multiple existing datasets from the literature. The proposed method shows high prediction performance when compared with four existing intrusive and non-intrusive SIP algorithms. This underlines the strong potential of deep learning for use in predicting speech intelligibility.

1 Introduction

Algorithms for Speech Intelligibility Prediction (SIP) typically attempt to predict the average intelligibility of noisy or degraded recordings of speech, to a group of average normal-hearing listeners. This can be an advantageous alternative to carrying out time consuming and expensive listening experiments involving many test subjects. Such algorithms were first studied in the telephone industry, with the aim of quantifying intelligibility of speech transmitted via telephone, without relying on listening experiments [1, 2]. This research resulted in the Articulation Index (AI), which is an objective scoring of intelligibility, between zero and one [1]. This scoring is computed as a weighted sum of contributions from a range of non-overlapping frequency bands [1]. The band-wise contributions are, in turn, based on the long-term Signal to Noise Ratio (SNR) within each band [1]. The resulting scoring of intelligibility has been shown to have a nearly monotonic relationship with measured intelligibility (measured in percentage of words understood correctly). It has, furthermore, been shown that the AI can be interpreted as an estimate of the information capacity of a noisy communication channel [3]. An updated and ANSI-standardized version of the AI is known as the Speech Intelligibility Index (SII) [4].

Since the introduction of the AI, a considerable amount of research has been carried out within SIP. This, more recent, work often aims to provide predictions in conditions where the AI and SII fall short. These include conditions with fluctuating interferers, e.g. [5–10], conditions where the speech signal has been non-linearly degraded or processed, e.g. [11, 12], conditions with

binaural listening, e.g. [7, 8, 13–18], and conditions with hearing impaired listeners, e.g. [19–22]. Furthermore, SIP algorithms can be classified according to the input signals required for making predictions. The AI and SII require access to clean speech and noise in separation (and therefore do not handle non-linearly degraded noisy speech) [1]. A second class of SIP algorithms makes predictions based on the degraded speech signal (i.e. that for which intelligibility is predicted), and either the clean speech in separation, e.g [10–12, 21, 23], or the clean noise in separation, e.g. [9, 18, 24]. Together, SIP algorithms which require knowledge of the clean speech signal or the clean noise signal, are known as *intrusive* SIP algorithms [22].

Recently, however, research has been increasingly directed towards *non-intrusive* SIP algorithms, which predict intelligibility by using only the degraded speech signal, e.g. [19, 20, 22, 25–31]. Such algorithms could be valuable for online assessment of speech intelligibility in signal processing devices such as hearing aids, or for other real-world applications where a clean signal cannot be obtained [22, 29–31].

One intrusive SIP algorithm has recently gained considerable popularity in the signal processing community: the Short-Time Objective Intelligibility (STOI) measure [12]. This very simple algorithm predicts intelligibility from clean and degraded speech, by averaging the sample correlation coefficient of short segments of clean and degraded envelopes across 15 one-third octave bands. The STOI measure has shown to correlate well with measured intelligibility in conditions including different additive noise sources [12], noise reduction processing [12, 32], hearing-aid and cochlear implant processing [22], and noisy speech transmitted via telephone [33]. Later work has shown the STOI measure to be closely related to an estimate of information transmission [23]. The STOI measure has been extended to make binaural predictions [34, 35], as well as to make non-intrusive [30], or partly non-intrusive [29, 31], predictions. An extension of the STOI measure, which aims to increase prediction accuracy for speech masked by fluctuating noise, is proposed in [10]. Another extension aims to increase prediction accuracy by weighting contributions of speech segments according to the information content of the speech [36]. A noise reduction algorithm, based on maximizing the STOI measure, is presented in [37].

The SIP algorithms discussed until this point are typically based on heuristics, as well as simple models of the human auditory system, and in some cases information theory. The algorithms gain trustworthiness primarily from repeated displays of accurate prediction performance across different conditions. Somewhat surprisingly, the SIP problem has only been sporadically investigated from a machine learning perspective [38–40]. The SIP problem has a clearly defined structure, with one or more well-defined input signals, and one output between zero and one. Thus, it can be considered as a regression problem. Furthermore, the recent success of deep

learning, within both Automatic Speech Recognition (ASR) [41] and speech enhancement [42, 43], suggests that similar approaches could be successful within SIP. The absence of such work may be primarily due to the lack of widely available datasets of degraded speech and corresponding measured intelligibility.

In this paper, we use Convolutional Neural Networks (CNNs) to non-intrusively predict the intelligibility of speech which has been degraded in different ways. When using neural networks for ASR, these typically include millions of parameters, and are trained using hundreds or thousands of hours of speech material [41]. We are not aware of the existence of such large databases of measured intelligibility, and this consequently rules out the possibility of training similarly large neural networks for SIP. Instead, this work has been guided by the following hypothesis:

> *SIP is a simple problem in comparison with ASR and speech enhancement, and can therefore be solved by a comparatively smaller neural network.*

Under this hypothesis, it should be possible to obtain good performance on the SIP task, without having to rely on massive quantities of training data. The hypothesis is based on the belief that speech intelligibility can be effectively assessed by the presence or absence of a rather small number of spectro-temporal patterns in a signal. This belief is motivated by the observation that existing non-intrusive SIP algorithms are able to predict intelligibility accurately across many conditions, by considering very simple modulation-domain structures [25, 26, 30]. In contrast, an ASR system must, in some form or another, store a detailed lexicon of all sounds present in speech, such as to allow for distinguishing and classifying these.

In order to make our work interpretable and comparable to existing non-intrusive and intrusive SIP algorithms, we have used a simple CNN structure, which is based on a small number of easy-to-visualize modulation features. In Section 2, we provide a detailed description of the structure of this network. In Section 3, we describe the speech database used for training, validation, and testing of the method. In Section 4, we evaluate the proposed measure, and compare it with existing non-intrusive and intrusive SIP algorithms. In Section 5 we discuss the implications of specific design choices in the proposed CNN architecture and the approach used for training it. Section 6 concludes upon our findings.

2 A CNN for Intelligibility Prediction

In this section we provide a detailed overview of the CNN architecture which we use to non-intrusively predict speech intelligibility. We specifically chose

a CNN structure because 1) this allows for handling input signals of varying length, and 2) the resulting convolution kernels can be visually inspected, and clearly reveal the spectro-temporal features used by the network.

Since we consider non-intrusive intelligibility prediction, only a degraded input signal, $y(t)$, is available. This is assumed to be a recording of at least several spoken sentences, which may be degraded by noise, reverberation, non-linear distortion, or essentially any other factor. The goal is to predict the fraction of words which are understandable to a normal-hearing listener. By a word being "understandable", we mean that the listener is able to repeat it, after having heard it. Thus, the output is a number in the range 0% to 100%.

The input signal, $y(t)$, is preprocessed before being presented to the aforementioned CNN. Specifically, this is done to lower computational demands. Furthermore, the preprocessing steps serve to make the resulting network independent of the overall level of the input speech. The preprocessing consists of steps which are highly similar to ones carried out in the STOI measure [12]. These steps result in a considerably more compact signal representation, and the success of the STOI measure suggests that they do not remove information which is crucially important to speech intelligibility.

2.1 Preprocessing

The input signals are first resampled to 10 kHz and periods without speech are removed by use of an ideal Voice Activity Detector (VAD), i.e. a VAD which makes use of the underlying clean speech signal. An ideal VAD is used to allow for meaningful performance evaluation. See Section 5 for a discussion of the necessity and implications of this. The VAD consists of two steps: 1) both the clean and degraded signals are segmented into 256-sample Hann-windowed segments, with an overlap of 50% between consecutive frames, and 2) the degraded signal is resynthesized, using only the frames with a clean speech power of more than -40 dB relative to the most powerful frame. These steps are identical to ones used in the STOI measure [12].

After resampling and removal of segments without speech, the signal is analyzed with a short-time Discrete Fourier Transformation (DFT), following the exact structure of the STOI measure [12]. This is done in 256-sample Hann-windowed segments which are zero-padded to 512 samples. The DFT coefficient corresponding to the kth frequency bin and the mth time frame is denoted $\hat{y}_{k,m}$.

Envelopes are then extracted in $Q = 15$ one-third octave bands, across the M time frames in the signal, in the same way as in the STOI measure [12]:

$$Y_{q,m} = \sqrt{\sum_{k=k_1(q)}^{k_2(q)} |\hat{y}_{k,m}|^2}, \tag{H.1}$$

for $m = 1, \ldots, M$ and $q = 1, \ldots, Q$, where q is the one-third octave band index, and $k_1(q)$ and $k_2(q)$ are, respectively, the lower and upper limits of the qth one-third octave band. The one-third octave bands have center frequencies spaced by one third octave, starting at 150 Hz.

The envelopes are mean- and variance normalized. This is also done in the STOI measure [12], but in a slightly different manner. We define the normalized envelope sample, $\bar{Y}_{q,m}$, by the two following steps:

$$\tilde{Y}_{q,m} = Y_{q,m} - \frac{1}{N} \sum_{m'=m-N+1}^{m} Y_{q,m'}, \tag{H.2}$$

for $m = N, \ldots, M$, and:

$$\bar{Y}_{q,m} = \frac{\tilde{Y}_{q,m}}{\sqrt{\frac{1}{N} \sum_{m'=m-N+1}^{m} \tilde{Y}_{q,m'}^2}}, \tag{H.3}$$

for $m = 2N - 1, \ldots, M$, where $\tilde{Y}_{q,m}$ is a zero-mean intermediate variable, and $\bar{Y}_{q,m}$ is the normalized envelope. As in the STOI measure, we use $N = 30$ envelope samples (corresponding to 384 ms) to estimate the mean and variance. The resulting normalized envelopes are defined for $Q = 15$ one-third octave bands, and for $L = M - 2N + 2$ time windows.

2.2 Network Architecture

The preprocessed input signal can be represented as a matrix of $L \times Q$ real numbers. The aim is to make a prediction of the intelligibility of the input signal in the range 0–100%. The specific CNN architecture, used to do so, is illustrated in Fig. H.1.

First, the preprocessed input signal is convolved with K kernels of dimension $N \times Q$ (marked (A) in Fig. H.1). This convolution is carried out only for shifts where the kernel is entirely contained in the preprocessed signal (this is sometimes referred to as "narrow" convolution). Furthermore, the convolution is carried out with a subsampling (or a stride) of 10, to limit computational demand. Thus, the output of this convolution is a tensor of dimension $(L - N + 1)/10 \times 1 \times K$.[1] A kernel-dependent constant is added to this tensor (marked (B) in Fig. H.1). The temporal length of the kernels has been chosen to $N = 30$ (or 384 ms), because a similar time scale was used with good results for both the STOI [12] and Non-Intrusive STOI (NI-STOI) [30] measures. The kernels were chosen to span all $Q = 15$ frequency bands, in order not to impose any constraints on the modeled structures across frequency. Furthermore, this design allows for easy visual inspection and interpretation

[1] More precisely the first dimension is $\lfloor (L - N + 1)/10 \rfloor$. We have left out the floor operator, $\lfloor \cdot \rfloor$, for notational convenience.

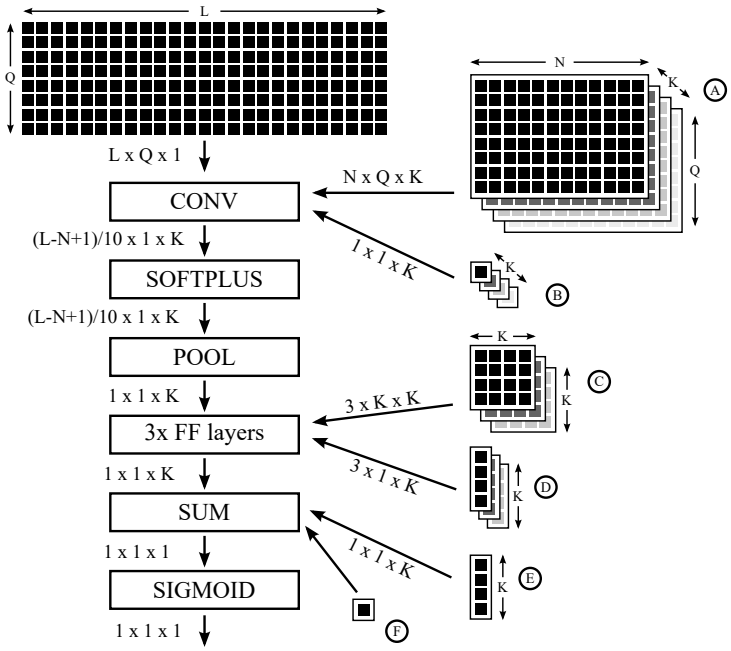


Fig. H.1: A block diagram of the applied network architecture. Network weights, indicated by (A)–(F), are found by stochastic gradient descent.

of the kernels. The result of the convolution is transformed in an entry-wise manner with a "*softplus*" non-linearity [44]:

$$f(z) = \log(1 + e^z). \tag{H.4}$$

Following this, pooling is carried out by averaging across the first dimension of the signal, yielding an output vector of dimension $1 \times 1 \times K$. This vector is passed through three conventional Feed-Forward (FF) layers (weights marked (C) in Fig. H.1), with K inputs and K outputs, each employing softplus non-linearities. A constant-vector is added before each non-linearity (marked (D) in Fig. H.1). A weighted summation of the the K outputs of the final FF layers is carried out (weights marked (E) and constant marked (F) in Fig. H.1), followed by a sigmoid non-linearity. The resulting output is a number in the range 0 to 1, corresponding to 0–100% predicted intelligibility. The system was implemented using Theano [45].

2.3 Interpretation of the Architecture

The proposed CNN architecture can be related to the structure of existing SIP algorithms. The majority of existing SIP algorithms assume that

additive contributions to intelligibility are supplied from different frequency bands [1, 4, 12] or modulation frequency bands [24, 46, 47]. Thus, contributions from several separate channels are computed and linearly combined, typically applying some numerical weighting of the importance of each channel [1, 48]. The contributions typically depend on either the SNR [1] or the correlation between clean and degraded envelopes [10–12, 23] within a band. For instance, the SII is computed as follows [4]:

$$\text{SII} = \sum_{i=1}^{n} I_i A_i, \tag{H.5}$$

where n is the number of frequency bands, I_i is the relative importance of the i'th band, and A_i is a measure of the speech fidelity in the ith band, which is, in turn, determined by transforming the SNR, in the ith band, with a compressive activation function. The resulting sum of contributions can then be transformed into a prediction of intelligibility (as a fraction of correctly understood words), by use of a function with an output in the range 0 to 1 (e.g. a logistic function [12] or a cumulative normal distribution [24]).

In the CNN architecture proposed in this paper, the K outputs of the convolution stage react to the presence of different sprectro-temporal patterns in the input signal. These excitations are non-linearly transformed and mixed in three FF layers. The outputs from the last FF layer are summed together using trained weights. The resulting sum is transformed into the range 0 to 1, by a sigmoid function. The last part of the CNN is similar to existing SIP algorithms, in that it computes a weighted sum of contributions, which is then transformed onto the interval 0 to 1 with a mapping function. These last two steps of this process correspond closely to the computation of e.g. the SII. However, instead of combining the SNR-based values A_i, the CNN architecture combines the outputs of the FF network into an index. This index is transformed into a prediction of intelligibility in percent by a sigmoid function. A similar approach is used to transform the SII into a prediction of intelligibility in percent [4]. The difference between the proposed CNN architecture and the SII is, then, that the proposed architecture non-intrusively estimates the K separate contributions, while the SII relies on the band-wise SNR, transformed by a compressive non-linearity. The proposed architecture does so by detecting the presence of a range of trained modulation templates (i.e. kernels), which may be indicators of speech, noise, or other factors which impact intelligibility. This in turn suggests that the combination of the convolutional stage and the FF network can be interpreted, in part, as an SNR estimator, followed by a non-linear transformation similar to that used in the SII. An advantage of the proposed method is that visual inspection of the kernels can give an understanding of which features the network associates with speech.

2.4 Training

The system is trained in a supervised manner on a database of audio files of different lengths, each with an associated measured intelligibility (see Section 3). Network weights (marked (A)–(F) in Fig. H.1) were found using stochastic gradient descent [49], while optimizing for the cross entropy [49]. Each gradient step is computed from a minibatch assembled from 5 seconds of audio, picked from a uniformly distributed random location in the signal, from each of five randomly picked audio files, i.e a total of 5×5 seconds of audio, and 5 corresponding values of measured intelligibility between 0% and 100%. We consider one epoch to constitute a single use of each value of measured intelligibility. Consequently, one epoch makes use of only 5 seconds of audio from each associated audio file.

Training commences with a learning rate of 0.01. Performance is evaluated on both the training set and the validation set once every 50 epochs (see Section 3 for details on the training and validation sets). At this point, the learning rate is decreased by 15% if 1) the current best training set performance has not been found within any of the last five performance evaluations, or 2) the current best validation set performance has not been found within any of the last 20 performance evaluations. Training is stopped when the learning rate becomes too small for the gradient steps to impact performance. Evaluation is performed using the weights which gave rise to the highest validation set performance.

The training process was regularized in two different ways. First, a regularization term, consisting of 10^{-5} times the sum of squares of all trained weights, was added to the objective function. Secondly, dropout, with a probability of 0.5 was used on the weights in the convolution stage. Dropout was not used in the remaining stages, as this was found to make the training process unstable. This is most likely because relatively narrow FF layers were used in practice (i.e. there were few nodes in each layer).

3 Data Material

To evaluate the proposed architecture, we assembled a dataset by combining data from a number of sources. This dataset was split into training, validation and testing sets in two different manners, as described below.

3.1 Sources of Data

We obtained measured intelligibility from four different listening experiments, previously described in the literature (see Table H.1):

- **$\mathbf{D_1}$:** (Described in [50].) Intelligibility was measured for noisy sentences processed with Ideal Time Frequency Segregation (ITFS). This

Table H.1: An overview of the four listening experiments.

ID	Collected by	# conditions
D_1	Kjems et al. [50]	114
D_2	Kjems et al. [50]	33
D_3	Jensen & Taal [10]	60
D_4	Studebaker & Sherbecoe [51]	318

was done for 1) four different noise types: Speech Shaped Noise (SSN), bottling factory hall noise, car noise, and café noise, 2) two different types of ideal binary masks, 3) eight different Relative Criterion (RC) values (parameter setting for the ITFS algorithms), and 4) three different SNR values. The measurements were carried out for 15 normal hearing subjects, using the Dantale II speech corpus [52]. In this work, we excluded the conditions with café noise, as this effectively consists of a single interfering talker. In conditions with a single interfering talker, a non-intrusive SIP algorithm must be supplied with additional information to correctly identify the target talker (as it could be any of the two talkers). We consider such conditions to be beyond the scope of this work.

- **D_2:** (Described in [50].) This experiment measured the intelligibility for speech masked by the same four types of noise, as used in dataset D_1, but without the application of ITFS. This was done for 15 normal hearing listeners with the Dantale II corpus, using an adaptive procedure for measuring the 20%– and 80% Speech Reception Thresholds (SRTs) [50][2]. The speech intelligibility was estimated at 11 SNRs, uniformly spaced from -20 dB SNR to 5 dB SNR, by fitting a logistic function to the measured SRTs, and interpolating [23]. The café noise conditions were removed from this dataset, for the same reason as for dataset D_1.

- **D_3:** (Described in [10].) Intelligibility was measured for Dantale II sentences masked by ten different types of noise, including SSN, unintelligible babble (1, 2, and 6 talkers), sinusoidally intensity modulated SSN (2, 4, 8, and 16 Hz), "machine gun" noise, and "destroyer operations room" noise (the two last noise types are from the NOISEX corpus [53]). Intelligibility was measured at six SNRs, uniformly spaced by 3 dB, and centered around the experimenters rough estimate of the 50% SRT [10]. The experiment was carried out for 12 normal hearing listeners.

- **D_4:** (Described in [51].) Intelligibility was measured for high- and low-pass filtered spoken words masked by SSN [51]. This was done using

[2]The X% SRT is the SNR at which the subject is able to correctly repeat X% of presented words.

recordings of the CID W-22 word lists [54]. Measurements were taken at 21 filter cutoffs and 10 SNRs, but some combinations were left out due to very low intelligibility. Eight normal hearing listeners participated in the study.

In all four listening experiments, the sound was delivered to the subject via headphones. For the first three datasets, we are in possession of the actual speech signals which were presented to the listeners. For dataset D_4, we recreated stimuli by high- or low-pass filtering speech and mixing it with SSN, following the description in [51].

The first three datasets were collected using the Danish Dantale II speech corpus [52] while D_4 was collected using the CID W-22 corpus [54]. It is worthwhile to notice that the Dantale II corpus includes only 50 unique words, while the CID W-22 corpus contains 200 unique words. By training on such small corpora, it is likely that the resulting CNN could be able to recognize individual words, or parts of words, and associate these with high levels of intelligibility. This is an unwanted behaviour, as a truly non-intrusive SIP algorithm should not be dependent on a particular underlying speech corpus. To avoid this problem, we recreated the signals of all four experiments using another Danish speech corpus: Akustiske Databaser For Dansk (ADFD)[3]. This includes a wide variety of Danish sentences spoken by more than 600 individuals of both genders. In this way, it was possible to ensure that a broad corpora of sentences spoken by different, non-overlapping, sets of talkers were used for training, validation, and testing. Speech material corresponding to at least ten sentences per condition was generated. This resulted in slightly less than 7 hours of audio in total (after voice activity detection).

The choices, involved in assembling the above described dataset, are further discussed in Section 5.

3.2 Training, Validation, and Testing Sets

The used datasets include measured intelligibility in a total of 525 conditions (see Table H.1). To facilitate an evaluation of the proposed CNN architecture, these conditions were split into training, validation, and testing sets. To do so, it is important to realize that varying levels of similarity exist between the conditions. For instance, intelligibility was measured for different SNRs for each noise/processing configuration, and each of these measurements are counted as one individual condition. If a system is trained with measurements for several SNR values in one noise/processing configuration, and tested on an unseen SNR in that same configuration, the task may be likened to a simple interpolation task. On the other hand, it may be a much more

[3]See `http://www.nb.no/sbfil/dok/nst_taledat_dk.pdf`.

difficult task to correctly predict the intelligibility for a noise/processing configuration which was not included in the training data. If the 525 conditions are split randomly, the evaluation may, thus, consist of tasks of somewhat varying difficulty. In particular, the results may not be representative of the performance which could be obtained with entirely novel types of noise and processing.

To alleviate this potential problem, we impose some structure on the generation of training, validation, and testing subsets. We investigate two different ways of doing so: 1) we perform the split such that, to the furthest extent possible, all noise/processing configurations are represented in all three subsets, but at different SNRs, and 2) we perform the split such that no noise/processing configuration is represented in more than one subset. The first method allows us to investigate the ability of the trained network to generalize to previously *unseen* SNRs in previously *seen* configurations of noise type and processing. The second method allows us to investigate the ability of the trained network to generalize to *unseen* configurations of noise type and processing.

In the first type of split (unseen SNRs), the data is split to use approximately four of six conditions for training and one of six conditions for each of validation and testing. Whenever this distribution cannot be obtained exactly, the remaining data points are distributed via fair lottery.

The second type of split (unseen noise/processing configurations) is performed slightly differently. We also use four out of six conditions for training, and one of six for each of validation and testing. However, since the different conditions may not be equally difficult to predict, this introduces a considerable source of variability, i.e. the performance on a particular testing set is likely to depend considerably on the specific noise/processing configurations contained in that set. Therefore, to ensure that all conditions are equally represented in the evaluation, we carry out the second type of split as Leave-N-Out Cross Validation (LNOCV). To do so, the noise/processing configurations were randomly split into six subsets, containing approximately the same number of conditions. Four of these subsets are used for training, one for validation, and one for testing. By rotating which subsets are used for training, validation and testing, we obtain six different splits, such that every data point is included exactly once in a validation set and once in a testing set. This ensures that predictions are made for all conditions in the available dataset.

4 Results

In this section we evaluate the proposed CNN, and compare it with four previously proposed non-intrusive and intrusive SIP algorithms. We sepa-

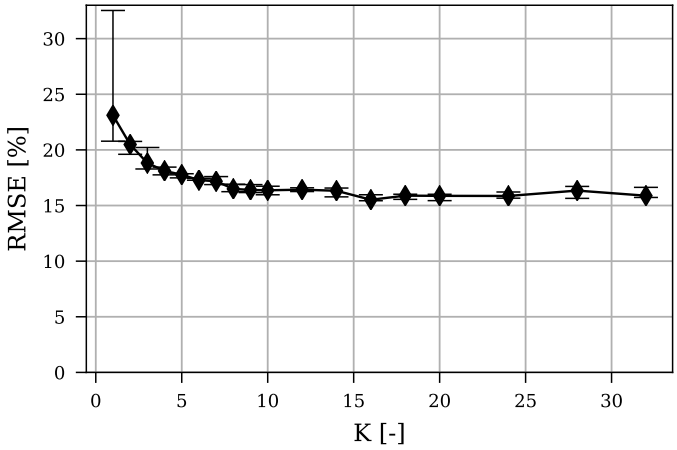


Fig. H.2: The median test RMSE vs K, when testing with unseen SNRs but previously seen noise/processing configurations. The error bars show 25th and 75th percentiles.

rately evaluate performance for the two different approaches for generating training, validation, and testing subsets (described in Section 3.2). Lastly, we investigate the properties of a specific instance of a trained network.

4.1 Testing with Unseen SNRs

We first consider performance when training with all noise/processing configurations, and testing with unseen SNRs (the first data splitting method described in Section 3.2). Because of the relatively small size of the dataset, random factors in the splitting procedure may affect performance. We therefore drew ten realizations of the split, and evaluated performance for each of these.

Fig. H.2 shows the relationship between median prediction performance, in terms of Root-Mean-Square Error (RMSE), and the value of K (i.e. the number of convolution kernels). The median is computed across the performance of ten trained networks; one for each dataset split. The figure shows performance improving for increasing K, until around $K = 16$, where performance plateaus. The error bars indicate that performance is very consistent across the different realizations of splits, except at low values of K. In absolute terms, the best RMSE is slightly below 16%. This suggests a considerable degree of correlation with measured intelligibility.

Based on the results of Fig. H.2, we use $K = 16$ for testing with unseen noise/processing configurations, as described in the following section.

4.2 Testing with Unseen Noise/Processing Configurations

We now consider performance for noise/processing configurations which were not used for training. To do so, we split the dataset by the second method described in Section 3.2, including the use of LNOCV. We generated ten sets of six splits, as described in Section 3.2. One CNN was trained for each resulting split, yielding ten different predictions for each condition (i.e. 60 CNNs were trained in total). The medians of these predictions, for each condition, are plotted against measured intelligibility in Fig. H.3a. This shows that accurate predictions are made for the majority of conditions. This, to a certain extent, suggests that the trained CNNs have learned fundamental features which govern the intelligibility of speech. Especially the conditions of dataset D_4 consistently appear to be very accurately predicted. Datasets D_1 and D_2 are also reasonably predicted. Dataset D_3, on the contrary, is notably less accurately predicted. This dataset contains speech in different types of modulated or strongly fluctuating noise types. Such conditions have proved difficult to predict by several existing SIP algorithms, including the SII and the STOI measure [6, 10].

The prediction bias (i.e. the average signed prediction error) of the proposed method, for datasets D_1, D_2, D_3 and D_4 is, respectively -0.04%, 12.1%, 3.64%, and -2.44%. Disregarding the bias for dataset D_2, which contains relatively few conditions, the bias is very small, i.e. there is little tendency to systematically over- or underestimate intelligibility from particular listening experiments. This can be taken as an indication that merging different datasets is not problematic in this context (we discuss this issue further in Section 5).

The ten largest prediction errors on Fig. H.3a are annotated and listed in Table H.2. Not surprisingly, these mainly consist of conditions from dataset D_3. These are conditions with noise types that contain speech-like modulations. More surprisingly, a large prediction error occurs for speech in SSN at an SNR of -5 dB (No. 7 in Table H.2, dataset D_3). From Fig. H.3, it can be seen that intelligibility, in this condition, is predicted to be around 40%, while it is measured to be around 90%. Similar conditions (i.e. speech masked by SSN) are included in dataset D_4. Here intelligibility has been measured at SNRs of -4 dB and -6 dB, to, respectively, 73.2% and 49.2% [51]. Both are considerably lower than the result from dataset D_3. This may be due to difference in the intelligibility of the Dantale II corpus used for collecting D_3 and the CID W-22 corpus used for collecting D_4.

Fig. H.3b shows predictions by four other SIP algorithms: the intrusive STOI [12] and Extended STOI (ESTOI) [10] measures, as well as the non-intrusive NI-STOI measure [30] and the Speech to Reverberation Modulation energy Ratio (SRMR) [25, 28]. To ensure a fair comparison, we preprocessed the input signals, for these algorithms, with exactly the same ideal VAD as

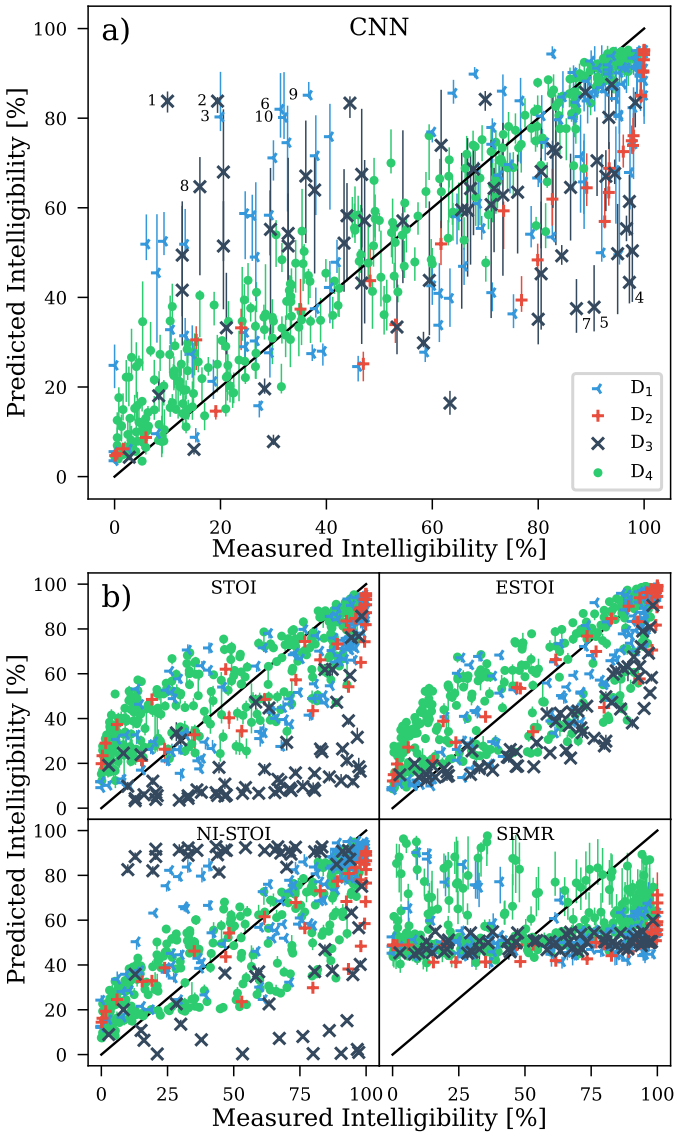


Fig. H.3: The median of predictions for the proposed CNN, and for four other SIP algorithms, plotted against corresponding measured intelligibility. The error bars show the 25th and 75th percentiles of predictions. Colors/symbols indicate which dataset each condition belongs to. For the proposed method, the ten conditions with the largest absolute prediction errors are numbered in descending order. Descriptions of these are given in Table H.2.

Table H.2: Conditions with largest absolute prediction errors.

No.	Error	Description
1	73.8%	D_3 • ICRA7: 6 spkr babble • SNR=–19dB SNR
2	64.4%	D_3 • ICRA7: 6 spkr babble • SNR=–16dB SNR
3	60.3%	D_1 • Car cabin • SNR=–60dB • TBM RC=12.7dB
4	–53.9%	D_3 • Mod. SSN f=16Hz • SNR=–10dB
5	–52.7%	D_3 • Mod. SSN f=16Hz • SNR=–13dB
6	50.7%	D_1 • Car cabin • SNR=–20.3dB • TBM RC=12.7dB
7	–49.7%	D_3 • ICRA1: SSN • SNR=–5dB
8	48.6%	D_3 • Mod. SSN f=8Hz • SNR=–25dB
9	48.5%	D_1 • Bottling factory hall noise • SNR=–60dB • IBM RC=-25.2dB
10	48.2%	D_1 • Car cabin • SNR=–23dB • TBM RC=12.7dB

Table H.3: Performance metrics for the five SIP algorithms.

SIP algorithm	RMSE	Kendall's Tau
CNN	17.67%	0.678
STOI	24.35%	0.550
ESTOI	19.74%	0.633
NI-STOI	23.38%	0.566
SRMR	34.41%	0.232

used in preprocessing for the proposed architecture. The outputs of these predictors were transformed with a logistic function [12]:

$$f(x) = \frac{1}{1 + \exp(ax + b)}. \tag{H.6}$$

The constants a and b were fitted to the training data[4], and predictions were carried out for the corresponding test set. This was done for all 60 dataset splits, yielding ten different predictions for each point.

The two intrusive SIP algorithms (the STOI and ESTOI measures) also show consistently good prediction performance, with the exception of a group of conditions, for which the STOI measure severely underestimates intelligibility. These are mainly conditions from dataset D_3, which involve modulated or otherwise fluctuating interferers. The ESTOI measure was developed specifically to cope better with such conditions [10]. It is entirely expected that these algorithms perform favorably in comparison with the proposed one, as they have a considerable advantage, in having the clean signal available. This is not the case for the two non-intrusive SIP algorithms

[4]The fitting was carried out using the `scipy.optimize.curve_fit`-function in SciPy [55].

included in the analysis (the NI-STOI measure and the SRMR). The NI-STOI measure generally predicts the trend accurately, but over- or underestimates intelligibility considerably for a number of conditions, particularly ones belonging to dataset D_3. The SRMR correlates poorly with measured intelligibility in the studied conditions.

A notable feature of Fig. H.3, is the fact that the measures in the STOI-family provide very consistent predictions across different training sets (i.e. the error bars are small), while the CNN-based approach shows considerable variability. This is certainly a consequence of the fact that the STOI-family measures have only two fitted parameters (a and b in (H.6)), while the CNN-based approach has many more. However, the variations are rather small in comparison with the magnitude of prediction errors, and are therefore not likely to influence performance strongly. Even more consistent predictions could possibly be obtained with a larger and more representative dataset.

Table H.3 lists the overall prediction performance of the five SIP algorithms in terms of RMSE and Kendall's tau. Both performance measures show the proposed method to perform similarly to or better than the four SIP algorithms used for comparison. This result should be considered in the light of the fact that the STOI and ESTOI measures gain a considerable advantage by having access to the clean speech signal. On the other hand, the CNN-based method is trained on conditions which, while not identical, may be similar to the ones used for testing. It is also notable that the non-intrusive NI-STOI measure performs slightly better than the original intrusive STOI measure.

4.3 Interpretation of Trained Network

Because of the simple structure of the used CNN and the small values of K, it is possible to effectively visualize the operation of the resulting network. To do this, we selected one particular network with $K = 10$, trained according to the procedure applied in Section 4.1. In this section, we illustrate the operation of this network in further detail.

The kernel weights of this network are plotted in Fig. H.4. Each kernel spans $Q = 15$ one-third octave bands, logarithmically spaced from 150 Hz to 3.8 KHz, and $N = 30$ time frames, corresponding to 384 ms. The convolutional part of the network can be interpreted as if looking for segments of signal with a time-frequency pattern similar to the kernels. The output of the convolutional stage is passed through three FF layers. It is therefore not directly possible to observe the effect of each kernel, on predicted intelligibility. A high degree of excitation for a particular kernel, may be interpreted, by the later stages of the network, as an indication of either higher intelligibility or of lower intelligibility (e.g. a kernel can function as a detector of intelligible speech, but also as a detector of noise). The kernels seen in Fig. H.4

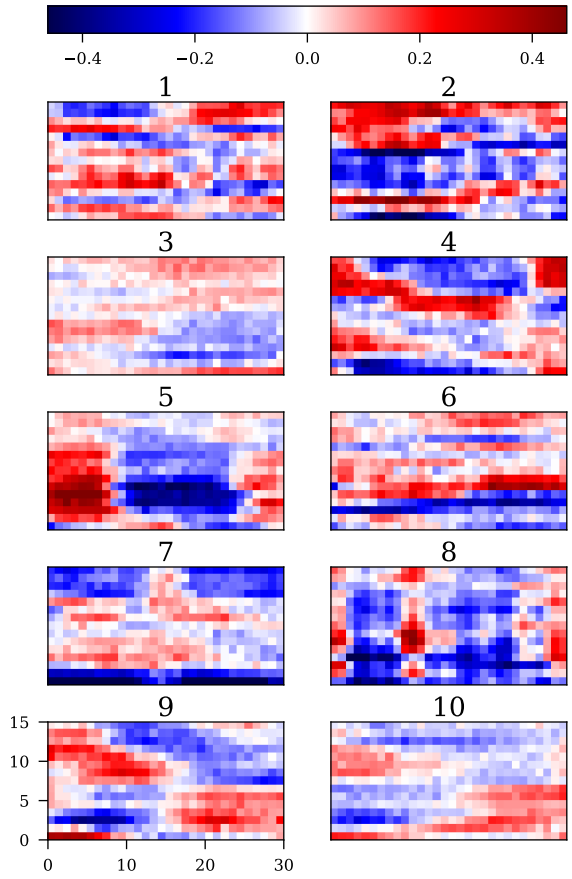


Fig. H.4: Plots of the kernel weights for a CNN with $K = 10$, trained according to the procedure described in Section 4.1.

clearly represent various types of spectro-temporal modulation structure. For instance, kernels 5 and 8 appear to encode temporal modulation at rates between 3 and 10 Hz. Kernel 5 could also be interpreted as a detector of short gaps. Kernels 3, 4, 9 and 10 appear to represent more complex temporal developments with the spectral distribution changing across time.

To further illustrate the operation of the network, Fig. H.5 shows several internal properties of the network, for an input signal consisting of a short segment of bottling factory hall noise followed by a short segment of speech and bottling factory hall noise at +10 dB SNR. Fig. H.5a shows a conventional spectrogram of the signal. Fig. H.5b shows the signal representation used as input to the network (i.e. the representation given by (H.3)). Fig. H.5c shows the ten outputs of the convolution stage (i.e. obtained by convolution with

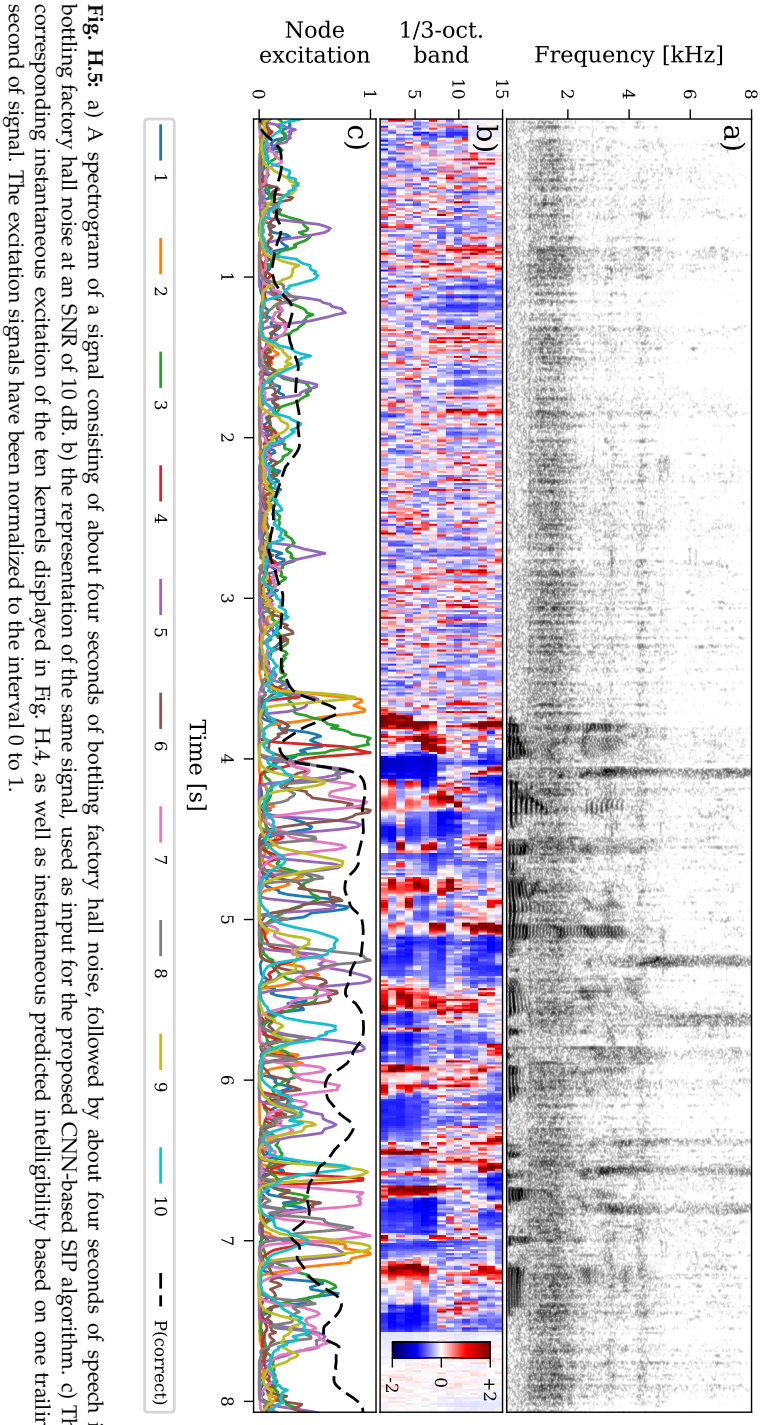


Fig. H.5: a) A spectrogram of a signal consisting of about four seconds of bottling factory hall noise, followed by about four seconds of speech in bottling factory hall noise at an SNR of 10 dB. b) the representation of the same signal, used as input for the proposed CNN-based SIP algorithm. c) The corresponding instantaneous excitation of the ten kernels displayed in Fig. H.4, as well as instantaneous predicted intelligibility based on one trailing second of signal. The excitation signals have been normalized to the interval 0 to 1.

the kernels displayed in Fig. H.4), as well as the predicted intelligibility of the signal, computed based on one trailing second of signal. From Fig. H.5b, it is evident, that the non-speech and speech segments contain very different modulation structures. Most notably, the non-speech segment appears to contain more high-frequency modulations, while the speech segment contains more low-frequency modulations. From Fig. H.5c, it is evident that the kernels are strongly excited by the speech part of the signal, and much less so by the non-speech part. Fig. H.5c furthermore suggests a considerable degree of independence between the kernels, i.e. the outputs of the kernels appear mutually uncorrelated. The individual kernels are excited somewhat sparsely during the speech part of the signal, and each of them at different points in time. This could indicate that the training has caused the kernels to be excited by different features in speech signals. The predicted speech intelligibility, shown in Fig. H.5c, is mostly low in the non-speech part of the signal, and becomes close to one in the speech part of the signal. This suggests that the network is able to distinguish speech from the noisy background. However, the strong fluctuations of the prediction also suggests that one second of audio is not enough to accurately predict intelligibility (one could argue that speech intelligibility is not even a meaningful concept for such a short signal).

5 Discussion

In this section we further discuss some design decisions concerning the proposed CNN architecture and the procedure used for training it. Specifically we discuss 1) the use a an ideal VAD in the preprocessing step for an otherwise non-intrusive method, and 2) the joining of data across multiple listening experiments and the associated substitution of underlying clean speech material.

5.1 The Use of an Ideal VAD

The preprocessing steps carried on the degraded speech inputs to the proposed CNN architecture (Section 2.1) involve the use of an ideal VAD. The use of an ideal VAD together with an otherwise non-intrusive method may seem somewhat inconsistent. However, we argue that the use of a VAD is necessary in order to ensure a consistent presence of speech in the signals used as input, at least for carrying out meaningful training and evaluation of the proposed method. To illustrate this point, two hypothetical signals are shown in Fig. H.6. Each signal consists of two tokens of speech and/or noise. Signal 1 consists of one token of clean speech (e.g. a sentence), and a token of clean noise without speech. The first token is easily intelligible,

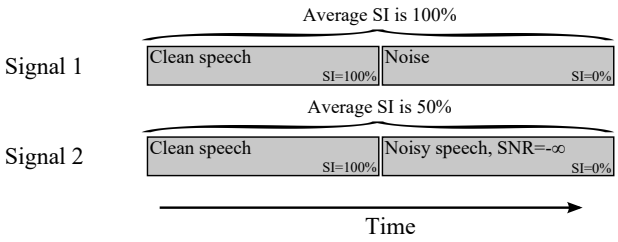


Fig. H.6: Two hypothetical signals, each composed of two consecutive segments. The first signal consists of a token of clean speech (e.g. a sentence), and a token of pure noise. The second signal consists of a token of clean speech, and a token of noisy speech, presented at an SNR which makes it indistinguishable from pure noise. The speech intelligibility of the first signal is 100%, as only the first token contains intelligible speech, and the second token contains no speech. The average intelligibility of the second signal is only 50%, as the first token contains intelligible speech, while the second token contains unintelligible speech.

and the second token contains no speech. Therefore all speech in Signal 1 is intelligible, i.e. the intelligibility is 100%. Signal 2 consists of one token of clean speech, followed by one token of noisy speech, presented at a SNR of $-\infty$ dB. In Signal 2, the first token of speech is intelligible, but the second one is not, i.e. the average intelligibility is 50%. However, there is no way, even in principle, to predict this difference from the degraded signals only, because signals 1 and 2 are, in fact, sample-wise identical: both signals contain one token of clean speech, and one token of noise. The only difference is whether the second token is considered to contain an underlying, unintelligible, speech signal, or not. This ambiguity arises for any noisy speech signal which contains speech pauses where only noise is present. We are not aware of any approach to resolve this ambiguity, that does not involve somehow standardizing the input signals. We therefore chose to standardize the input signals by ensuring that there are no segments with only noise. We remark that it is, of course, not fair to make performance comparisons between algorithms that do make use of an ideal VAD, and algorithms that do not. In the evaluation part of this work, we therefore used exactly the same VAD as a preprocessing step for all compared methods. For practical uses of non-intrusive SIP algorithms, this ambiguity has to be resolved in some other way (that does not assume knowledge of a clean signal). One possibility is to use a non-ideal VAD, which does not require knowledge of the clean signal (see e.g. [56]). Another approach could be to attempt estimation of a slightly different target than intelligibility in percent, such as the number of intelligible words or syllables per second of signal, or the rate of transferred information [3, 23]. Such targets do not suffer from similar ambiguities, as they are not dependent on the total number of words present in the underlying clean speech signal. It may even be possible to use the proposed measure in this

manner, by training with the use of an ideal VAD, and applying the resulting CNN with no VAD at all. In this case the output can be interpreted as a prediction of speech intelligibility under the assumption that the degraded signal includes a consistent presence of underlying clean speech. I.e. in practical applications, it may not be necessary to distinguish between situations where intelligibility is 0% because there is no speech, and situations where intelligibility is 0% because of severe degradations.

5.2 Joining Data Across Different Listening Experiments

The process of changing the underlying speech material in audio from listening experiments, as described in Section 3.1, introduces the assumption that the original target speech, used in the listening experiments, can be substituted by an alternative corpus of speech, and still be meaningfully used for predicting the intelligibility of the original speech signal. For this to be the case, the used CNN architecture should respond similarly to the original degraded speech and the degraded speech with substituted target speech. Because the employed architecture uses kernels with a temporal length of 386 ms, we assume it to be mainly sensitive to features on this time-scale, i.e. individual phonemes and transitions between these. The substitution should therefore be unproblematic, provided that the distribution of phonetic structures is similar across the original and substituted corpora. Since Dantale II and the ADFD corpora both contain common Danish sentences, we consider this to be a reasonable assumption for datasets D_1, D_2 and D_3. We were unable to obtain the used recordings of the CID W-22 word lists used in dataset D_4, and it is therefore more difficult to assess whether the substitution of speech corpora is justifiable in this case. Notably, the CID W-22 corpus contains English words rather than full Danish sentences. While we do not believe the difference between English and Danish to be highly important, it has previously been suggested that intelligibility is lower, when individual words are presented without a context [57]. This could possibly lead to some degree of discrepancy in the results.

The above discussion suggests an even broader question: is it sensible to merge results from different listening experiments, carried out with different subjects, different equipment, in different conditions, using different speech corpora? The answer, of course, depends on the magnitude of differences between the listening experiments, and the required quality of the merged dataset. All four datasets, considered in this work, have been collected with normal hearing subjects, presented with diotic degraded speech via headphones. Unless any of the equipment used in collecting these datasets have strongly impacted the measured results, it should be reasonable to assume that results can be compared across such studies. We believe the main difference to be the, already mentioned, difference between the used speech

corpora. The decision of whether to merge datasets is then a trade-off of quality against quantity. In this study we deemed it necessary to collect a considerable amount of data, in order to justify the training of networks with thousands of parameters. Thus, in contrast to many existing studies on SIP, we have placed a slightly larger focus on the quantity of data, believing this to cause only an insignificant reduction in the quality of the used data.

The results presented in Section 4 do not appear to suggest that the joining of multiple datasets has led to any severe issues in the present study.

6 Conclusion

We have proposed a Convolutional Neural Network (CNN) architecture for use in Speech Intelligibility Prediction (SIP). The architecture is designed with a specific focus on being interpretable and structurally comparable to existing SIP algorithms. To evaluate the performance of the architecture, we collected a dataset of measured intelligibility by combining the results of four listening experiments from the literature. The performance of the proposed method was shown to be similar to or higher than that of four existing intrusive and non-intrusive SIP algorithms. By investigating the weights fitted in a specific instance of the proposed architecture, it was possible to visualize how the resulting network detects speech-like modulations, and uses these to assess the degree of intelligibility in a degraded speech signal.

7 Acknowledgments

The authors would like to thank Morten Kolbæk for providing a reorganized version of the "Akustiske Databaser for Dansk" corpus, as well as for several helpful discussions. The work was funded by the Oticon Foundation and the Danish Innovation Foundation.

References

[1] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, no. 1, pp. 90–119, Jan. 1947.

[2] H. Fletcher and J. C. Steinberg, "Articulation testing methods," *Bell System Technical Journal*, vol. 8, no. 4, pp. 806–854, Oct. 1929.

[3] J. B. Allen, "The articulation index is a shannon channel capacity," in *Auditory Signal Processing: Physiology, Psychoacoustics and Models*, D. Pressnitzer, A. de Cheveigne, S. McAdams, and L. Collet, Eds. Springer Verlag, 2004, pp. 314–320.

[4] "Methods for calculation of the speech intelligibility index," American National Standards Institute, New York, United States, Standard, 1997.

[5] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, Apr. 2005.

[6] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 3988–3997, Dec. 2006.

[7] R. Beutelmann, T. Brand, and B. Kollmeier, "Revision, extension and evaluation of a binaural speech intelligibility model," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2479–2497, Dec. 2010.

[8] R. Wan, N. I. Durlach, and H. S. Colburn, "Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 768—-776, Aug. 2014.

[9] S. Jørgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *J. Acoust. Soc. Am.*, vol. 134, no. 1, pp. 436–446, Jul. 2013.

[10] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[11] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, Apr. 2005.

[12] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[13] H. vom Hövel, "Zur bedeutung der übertragungseigenschaften des aussenohrs sowie des binauralen hörsystems bei gestörter sprachübertragung," Ph.D. dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, 1984.

[14] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 120, no. 1, pp. 331–342, Apr. 2006.

[15] R. Wan, N. I. Durlach, and H. S. Colburn, "Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers," *J. Acoust. Soc. Am.*, vol. 128, no. 6, pp. 3678–3690, Sep. 2010.

[16] M. Lavandier and J. F. Culling, "Prediction of binaural speech intelligibility against noise in rooms," *J. Acoust. Soc. Am.*, vol. 127, no. 1, pp. 387–399, Jan. 2010.

[17] S. Jelfs, J. F. Culling, and M. Lavandier, "Revision and validation of a binaural model for speech intelligibility in noise," *Hearing Research*, vol. 275, no. 1–2, pp. 96–104, May 2011.

[18] A. Chabot-Leclerc, E. N. MacDonald, and T. Dau, "Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain," *J. Acoust. Soc. Am.*, vol. 140, no. 1, pp. 192–205, Jul. 2016.

[19] T. H. Falk, S. Cosentino, J. Santos, D. Suelzle, and V. Parsa, "Non-intrusive objective speech quality and intelligibility prediction for hearing instruments in complex listening environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vancouver, Canada: IEEE, May 2013.

[20] J. F. Santos and T. H. Falk, "Updating the SRMR-CI metric for improved intelligibility prediction for cochlear implant users," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 2197–2206, Dec. 2014.

[21] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, Jul. 2014.

[22] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[23] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 430–440, Feb. 2014.

[24] S. Jørgensen and T. Dau, "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," *J. Acoust. Soc. Am.*, vol. 130, no. 3, pp. 1475–1487, Sep. 2011.

[25] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE*

Trans. Audio, Speech, Language Process., vol. 18, no. 7, pp. 1766–1774, Sep. 2010.

[26] F. Chen, O. Hazrati, and P. C. Loizou, "Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure," *Biomedical Signal Processing and Control*, vol. 8, pp. 311–314, Dec. 2013.

[27] S. Cosentino, T. Marquardt, D. McAlpine, J. F. Culling, and T. H. Falk, "A model that predicts the binaural advantage to speech intelligibility from the mixed target and interferer signals," *J. Acoust. Soc. Am.*, vol. 135, no. 2, pp. 796–807, Feb. 2014.

[28] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*. Juan les Pins, France: IEEE, Sep. 2014, pp. 55–59.

[29] C. Sørensen, J. B. Boldt, F. Gran, and M. G. Christensen, "Semi-non-intrusive objective intelligibility measure using spatial filtering in hearing aids," in *The European Signal Processing Conference (EUSIPCO)*. Budapest, Hungary: EURASIP, Aug. 2016, pp. 1358–1362.

[30] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A non-intrusive short-time objective intelligibility measure," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, United States: IEEE, Mar. 2017, pp. 5085–5089.

[31] C. Sørensen, A. Xenaki, J. B. Boldt, and M. G. Christensen, "Pitch-based non-intrusive objective intelligibility prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, United States: IEEE, Mar. 2017, pp. 386–390.

[32] K. Smeds, A. Leijon, F. Wolters, A. Hammarstedt, S. Båsjö, and S. Hertzman, "Comparison of predictive measures of speech recognition after noise reduction processing," *J. Acoust. Soc. Am.*, vol. 136, no. 3, pp. 1363—-1374, Sep. 2014.

[33] S. Jørgensen, J. Cubick, and T. Dau, "Speech intelligibility evaluation for mobile phones," *Acta Acustica United with Acustica*, vol. 101, pp. 1016–1025, 2015.

[34] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "A binaural short time objective intelligibility measure for noisy and enhanced speech," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015.

[35] ——, "Predicting the intelligibility of noisy and non-linearly processed binaural speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 1908–1920, 2016.

[36] L. Lightburn and M. Brookes, "A weighted STOI intelligibility metric based on mutual information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China: IEEE, Mar. 2016, pp. 5365–5369.

[37] ——, "SOBM - a binary mask for noisy speech that optimizes an objective intelligibility measure," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Melbourne, Australia: IEEE, Sep. 2015.

[38] M. I. Mandel, "Learning an intelligibility map of individual utterances," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, United States: IEEE, Oct. 2013.

[39] ——, "Measuring time-frequency importance functions of speech with bubble noise," *J. Acoust. Soc. Am.*, vol. 140, no. 4, pp. 2542–2553, Oct. 2016.

[40] A. H. Andersen, E. Schoenmaker, and S. van de Par, "Speech intelligibility prediction as a classification problem," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. Vietri sul Mare, Italy: IEEE, Sep. 2016.

[41] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Oct. 2012.

[42] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 3029–3038, Oct. 2013.

[43] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE Trans. Audio, Speech, Language Process.*, vol. 25, no. 1, pp. 153–167, 2017.

[44] C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia, "Incorporating second-order functional knowledge for better option pricing," in *Neural Information Processing Systems (NIPS)*. Vancouver, Canada: MIT Press, Oct. 2001, pp. 1369–1376.

[45] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016. [Online]. Available: http://arxiv.org/abs/1605.02688

[46] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[47] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Communication*, vol. 41, pp. 331–348, 2003.

[48] J. M. Kates, "Improved estimation of frequency importance functions," *J. Acoust. Soc. Am.*, vol. 134, no. 5, pp. EL459–EL464, Nov. 2013.

[49] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2017.

[50] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.

[51] G. A. Studebaker and R. L. Sherbecoe, "Frequency-importance and transfer functions for recorded CID W-22 word lists," *Journal of Speech and Hearing Research*, vol. 34, pp. 427–438, Apr. 1991.

[52] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *International Journal of Audiology*, vol. 42, no. 1, pp. 10–17, Jan. 2003.

[53] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[54] I. J. Hirsh, H. Davis, S. R. Silverman, E. G. Reynolds, E. Eldert, and R. W. Benson, "Development of materials for speech audiometry," *The Journal of Speech and Hearing Disorders*, vol. 17, no. 3, pp. 321–337, Sep. 1952.

[55] E. Jones, T. Oliphant, P. Peterson, and others, "SciPy: Open source scientific tools for Python," Retrieved from `https://www.scipy.org` on 28/07-2017, 2001–. [Online]. Available: http://www.scipy.org/

[56] J. Ramirez, J. M. Gorriz, and J. C. Segura, "Voice activity detection. fundamentals and speech recognition system robustness," in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds. I-Tech Education and Publishing, 2007, pp. 1–22.

[57] G. A. Miller, G. A. Heise, and W. Lichten, "The intelligibility of speech as a function of the context of the test materials," *Journal of Experimental Psychology*, vol. 41, no. 5, pp. 329–335, May 1951.

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-050-0

AALBORG UNIVERSITY PRESS